PCT WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification 5 :
C12N 15/29, C07K 13/00
C12N 1/21, 1/19

A1

(11) International Publication Number: WO 91/19801

(43) International Publication Date: 26 December 1991 (26.12.91)

(21) International Application Number: PCT/GB91/00914

(22) International Filing Date: 7 June 1991 (07.06.91)

(30) Priority data:
9013016.2 11 June 1990 (11.06.90) GB

(71) Applicant *(for all designated States except US)*: MARS UK LIMITED [GB/GB]; 3D Dundee Road, Slough, Berkshire SL1 4JS (GB).

(72) Inventors; and
(75) Inventors/Applicants *(for US only)* : SPENCER, Margaret, Elizabeth [GB/GB]; 24 Totley Brook Road, Sheffield S17 3QS (GB). HODGE, Rachel [GB/GB]; 16 Mount Avenue, Leicester LE5 3RN (GB). DEAKIN, Edward, Alfred [GB/GB]; The Oakes, Oakes Park, Norton, Sheffield S8 8BA (GB). ASHTON, Sean [GB/GB]; 5 Chamberlain Court, Burncross, Chapeltown, Sheffield S30 4ZU (GB).

(74) Agents: SHEARD, Andrew, Gregory et al.; Kilburn & Strode, 30 John Street, London WC1N 2DD (GB).

(81) Designated States: AT (European patent), AU, BE (European patent), BR, CA, CH (European patent), DE (European patent), DK (European patent), ES (European patent), FI, FR (European patent), GB, GB (European patent), GR (European patent), HU, IT (European patent), JP, KR, LU (European patent), NL (European patent), NO, PL, RO, SE (European patent), SU, US.

Published
With international search report.
Before the expiration of the time limit for amending the claims and to be republished in the event of the receipt of amendments.

(54) Title: RECOMBINANT 47 AND 31kD COCOA PROTEINS AND PRECURSOR

(57) Abstract

47 kD and 31 kD proteins, and their 67 kD expression precursor, believed to be the source of peptide flavour precursors in cocoa (*Theobroma cacao*) have been identified. Genes coding for them have been probed, identified and sequenced, and recombinant proteins have been synthesised.

FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AT	Austria	**ES**	Spain	**MG**	Madagascar
AU	Australia	**FI**	Finland	**ML**	Mali
BB	Barbados	**FR**	France	**MN**	Mongolia
BE	Belgium	**GA**	Gabon	**MR**	Mauritania
BF	Burkina Faso	**GB**	United Kingdom	**MW**	Malawi
BG	Bulgaria	**GN**	Guinea	**NL**	Netherlands
BJ	Benin	**GR**	Greece	**NO**	Norway
BR	Brazil	**HU**	Hungary	**PL**	Poland
CA	Canada	**IT**	Italy	**RO**	Romania
CF	Central African Republic	**JP**	Japan	**SD**	Sudan
CG	Congo	**KP**	Democratic People's Republic of Korea	**SE**	Sweden
CH	Switzerland			**SN**	Senegal
CI	Côte d'Ivoire	**KR**	Republic of Korea	**SU**	Soviet Union
CM	Cameroon	**LI**	Liechtenstein	**TD**	Chad
CS	Czechoslovakia	**LK**	Sri Lanka	**TG**	Togo
DE	Germany	**LU**	Luxembourg	**US**	United States of America
DK	Denmark	**MC**	Monaco		

RECOMBINANT 47 AND 31kD COCOA PROTEINS AND PRECURSOR

This invention relates to proteins and nucleic acids derived from or otherwise related to cocoa.

The beans of the cocoa plant (*Theobroma cacao*) are the raw material for cocoa, chocolate and natural cocoa and chocolate flavouring. As described by Rohan ("Processing of Raw Cocoa for the Market", FAO/UN (1963)), raw cocoa beans are extracted from the harvested cocoa pod, from which the placenta is normally removed, the beans are then "fermented" for a period of days, during which the beans are killed and a purple pigment is released from the cotyledons. During fermentation "unknown" compounds are formed which on roasting give rise to characteristic cocoa flavour. Rohan suggests that polyphenols and theobromine are implicated in the flavour precursor formation. After fermentation, the beans are dried, during which time the characteristic brown pigment forms, and they are then stored and shipped.

Biehl *et al*, 1982 investigated proteolysis during anaerobic cocoa seed incubation and identified 26kD and 44kD proteins which accumulated during seed ripening and degraded during germination. Biehl asserted that there were storage proteins and suggested that they may give rise to flavour-specific peptides.

Fritz *et al*, 1985 identified polypeptides of 20kD and 28kD appearing in the cytoplasmic fraction of cocoa seed extracts at about 100 days after pollination. It appears that the 20kD protein is thought to have glyceryl acyltransferase activity.

In spite of the uncertainties in the art, as summarised above, proteins apparently responsible for flavour production in cocoa beans have now been identified. Further, it has been discovered that, in spite of Fritz's caution that "cocoa seed

mRNA levels are notably low compared to other plants" (*loc. cit.*), it is possible to apply the techniques of recombinant DNA techniques to the production of such proteins.

According to a first aspect of the invention, there is provided a 67kD protein of *Theobroma cacao*, or a fragment thereof.

The 67kD protein appears to be a primary translation product of interest in proteins involved in flavour production in cocoa. The 67kD protein may be processed *in vivo* to form 47kD and 31kD polypeptides.

According to a second aspect of the invention, there is provided a 47kD protein of *Th. cacao*, or a fragment thereof.

According to a third aspect of the invention, there is provided a 31kD protein of *Th. cacao* or a fragment thereof.

The term "fragment" as used herein and as applied to proteins or peptides indicates a sufficient number of amino acid residues are present for the fragment to be useful. Typically, at least four, five, six or even at least 10 or 20 amino acids may be present in a fragment. Useful fragments include those which are the same as or similar or equivalent to those naturally produced during the fermentation phase of cocoa bean processing. It is believed that such fragments take part in Maillard reactions during roasting, to form at least some of the essential flavour components of cocoa.

Proteins in accordance with the invention may be synthetic; they may be chemically synthesised or, preferably, produced by recombinant DNA techniques. Proteins produced by such techniques can therefore be termed "recombinant proteins". Recombinant proteins may be glycosylated or non-glycosylated: non-glycosylated proteins will result from prokaryotic expression systems.

Theobroma cacao has two primary subspecies, *Th. cacao cacao* and *Th. cacao sphaerocarpum*. While proteins in accordance with the invention may be derived from these subspecies. the invention is not limited solely to these subspecies. For example, many cocoa varieties are hybrids between different species; an example of such a hybrid is the trinitario variety.

The invention also relates to nucleic acid, particularly DNA, coding for the proteins referred to above (whether the primary translation products, the processed proteins or fragments). The invention therefore also provides, in further aspects:

- nucleic acid coding for a 67kD protein of *Th. cacao*, or for a fragment thereof;

- nucleic acid coding for a 47kD protein of *Th. cacao*, or for a fragment thereof;

- nucleic acid coding for a 31kD protein of *Th. cacao*, or for a fragment thereof;

Included in the invention is nucleic acid which is degenerate for the wild type protein and which codes for conservative or other non-deleterious mutants. Nucleic acid which hybridises to the wild type material is also included.

Nucleic acid within the scope of the invention will generally be recombinant nucleic acid and may be in isolated form. Frequently, nucleic acid in accordance with the invention will be incorporated into a vector (whether an expression vector or otherwise) such as a plasmid. Suitable expression vectors will contain an appropriate promoter, depending on the intended expression host. For yeast, an appropriate promoter is the yeast pyruvate kinase (PK) promoter: for bacteria an appropriate promoter is a strong lambda promoter.

Expression may be secreted or non-secreted. Secreted expression is preferred, particularly in eukaryotic expression systems; an appropriate signal sequence may be present for this purpose. Signal sequences derived from the expression host (such as that from the yeast alpha-factor in the case of yeast) may be more appropriate than native cocoa signal sequences.

The invention further relates to host cells comprising nucleic acid as described above. Genetic manipulation may for preference take place in prokaryotes. Expression will for preference take place in a food-approved host. The yeast *Saccharomyces cerevisiae* is particularly preferred.

The invention also relates to processes for preparing nucleic acid and protein as described above by nucleic acid replication and expression, respectively.

cDNA in accordance with the invention may be useful not only for obtaining protein expression but also for Restriction Fragment Length Polymorphism (RFLP) studies. In such studies, detectably labelled cDNA (eg radiolabelled) is prepared. DNA of a cultivar under analysis is then prepared and digested with restriction enzymes. Southern blotting with the labelled cDNA may then enable genetic correlations to be made between cultivars. Phenotypic correlations may then be deduced.

The invention will now be illustrated by the following non-limiting examples. The examples refer to the accompanying drawings, in which:

Figure 1 shows a map of the coding region of the 67kD protein, together with the inter-relationship of plasmids pMS600, pMS700 and pMS800, from which sequence data were obtained:

Figure 2 shows the complete nucleotide sequence of cDNA coding for the 67kD protein and the deduced amino acid sequence;

Figure 3 shows the amino acid sequence referred to in Figure 2;

Figure 4 shows the relationship between the 67kD protein and seed storage proteins from other plants;

Figure 5 shows a map of plasmid pJLA502;

Figure 6 shows schematically the formation of plasmid pMS900;

Figure 7 shows two yeast expression vectors useful in the present invention; vector A is designed for internal expression and vector B is designed for secreted expression;

Figure 8a shows, in relation to vector A, part of the yeast pyruvate kinase gene showing the vector A cloning site, and the use of *Hin-Nco* linkers to splice in the heterologous gene;

Figure 8b shows, in relation to vector B, part of the yeast alpha-factor signal sequence showing the vector B cloning site, and the use of *Hin-Nco* linkers to create an in-phase fusion;

Figure 9a shows how plasmid pMS900 can be manipulated to produce plasmids pMS901, pMS903, pMS907, pMS908, pMS911, pMS912 and pMS914;

Figure 9b shows how plasmid pMS903 can be manipulated to produce plasmids pMS904, pMS905, pMS906, pMS909 and pMS916;

Figure 10 shows maps of plasmids pMS908, pMS914, pMS912, pMS906, pMS916 and pMS910;

Figure 11 shows the construction of a plasmid to express the 67kD protein by means of the AOX promoter on an integrated vector in *Hansenula polymorpha*; and

Figure 12 shows the construction of a plasmid to express the 67kD protein by means of the AOX promoter in conjunction with the yeast α-factor secretory signal on an integrated vector in *Hansenula polymorpha*.

EXAMPLES

Example 1

Identification of the Major Seed Proteins

It is not practicable to extract proteins directly from cocoa beans due to the high fat and polyphenol contents, and proteins were, therefore, extracted from acetone powders made as follows. Mature beans from cocoa of West African origin (*Theobroma cacao amelonada*) were lyophilised and ground roughly in a pestle and mortar. Lipids were extracted by Soxhlet extraction with diethyl ether for two periods of four hours, the beans being dried and further ground between extractions. Polyphenols and pigments were then removed by several extractions with 80% acetone, 0.1% thioglycollic acid. After extraction the resulting paste was dried under vacuum and ground to a fine powder.

Total proteins were solubilised by grinding the powder with extraction buffer (0.05 M sodium phosphate, pH 7.2; 0.01 M 2-mercaptoethanol; 1% SDS) in a hand-held homogeniser, at 5mg/ml. The suspension was heated at 95°C for 5 minutes, and centrifuged at 18 K for 20 minutes to remove insoluble material. The resulting clear supernatant contained about 1 mg/ml total protein. Electrophoresis of 25 μl on an SDS-PAGE gel (Laemmli, 1970) gave three major bands, two of which were at 47 kD and 31 kD, comprising over 60% of the total proteins. The 47kD and 31kD proteins are presumed to be the polypeptide subunits of major storage proteins.

Characteristics of the Storage Polypeptides

The solubility characteristics of the 47 kD and 31 kD polypeptides were roughly defined by one or two quick experiments. Dialysis of the polypeptide solution against SDS-free extraction buffer rendered the 47 kD and 31 kD polypeptides insoluble, as judged by their ability to pass through a 0.22 micron membrane. Fast Protein Liquid Chromatography (FPLC) analysis also showed that the 47 kD and 31 kD polypeptides were highly associated after extraction with McIlvaines buffer pH 6.8 (0.2 M disodium hydrogen phosphate titrated with 0.1 M citric acid). The 47 kD and 31 kD polypeptides are globulins on the basis on their solubility.

Purification of the 47 kD and 31 kD polypeptides

The 47 kD and 31 kD polypeptides were purified by two rounds of gel filtration on a SUPEROSE-12 column of the PHARMACIA Fast Protein Liquid Chromatography system (FPLC), or by electroelution of bands after preparative electrophoresis. (The words SUPEROSE and PHARMACIA are trade marks.) Concentrated protein extracts were made from 50 mg acetone powder per ml of extraction buffer, and 1-2 ml loaded onto 2 mm thick SDS-PAGE gels poured without a comb. After electrophoresis the gel was surface stained in aqueous Coomassie Blue, and the 47 kD and 31 kD bands cut out with a scalpel. Gel slices were electroeluted into dialysis bags in electrophoresis running buffer at 15 V for 24 hours, and the dialysate dialysed further against 0.1% SDS. Samples could be concentrated by lyophilisation.

Example 2

Amino-acid Sequence Data from Proteins

Protein samples (about 10 μg) were subjected to conventional N-terminal amino-acid sequencing. The 47 kD and 31 kD polypeptides were N-terminally blocked, so cyanogen bromide peptides of the 47 kD and 31 kD peptides were

prepared, and some amino-acid sequence was derived from these. Cyanogen bromide cleaves polypeptide chains at methionine residues, and thus cleaved the 47 kD and 31 kD polypeptides gave rise to 24 kD and 17 kD peptides. In addition the 47 kD polypeptide gave a 20 kD peptide. The 24 kD and 17 kD peptides had the same 9 N-terminal amino-acid residues. This fact, combined with the obvious one that the 31 kD could not contain both peptides consecutively, suggested that the 24 kD peptide arose for a partial digest, where full digestion would yield the 17 kD peptide. The other striking conclusion is that the 47 kD and 31 kD proteins are related, and the 31 kD could be a further processed form of the 47 kD. The 9 amino-acid sequence was used to construct an oligonucleotide probe for the 47 kD/31 kD gene(s).

Example 3

Raising Antibodies to the 47 kD and 31 kD Polypeptides

Polyclonal antibodies were prepared using the methodology of Catty and Raykundalia (1988). The serum was aliquoted into 1 ml fractions and stored at -20°C.

Characterising Antibodies to the 47 kD and 31 kD Polypeptides

Serum was immediately characterised using the Ochterloney double-diffusion technique, whereby antigen and antibody are allowed to diffuse towards one another from wells cut in agarose in borate-saline buffer. Precipitin lines are formed where the two interact if the antibody 'recognises' the antigen. This test showed that antibodies to both antigens had been formed, and furthermore that extensive cross-reaction took place between the 47 kD and 31 kD polypeptides and their respective antibodies. This is further indication that the 47 kD and 31 kD polypeptides are closely related, as suggested by their cyanogen bromide cleavage patterns.

The gamma-globulin fraction of the serum was partially purified by precipitation with 50% ammonium sulphate, solubilisation in phosphate-buffered saline (PBS) and chromatography on a DE 52 cellulose ion-exchange column as described by Hill, 1984. Fractions containing gamma-globulin were monitored at 280 nm (OD_{280} of 1.4 is equivalent to 1 mg/ml gamma-globulin) and stored at -20°C.

The effective titre of the antibodies was measured using an enzyme-linked immunosorbant assay (ELISA). The wells of a polystyrene microtitre plate were coated with antigen (10-1000 ng) overnight at 4°C in carbonate coating buffer. Wells were washed in PBS-Tween and the test gamma globulin added at concentrations of 10, 1 and 0.1 μg/ml (approximately 1:100, 1:1000 and 1:10,000 dilutions). The diluent was PBS-Tween containing 2% polyvinyl pyrrolidone (PVP) and 0.2% BSA. Controls were preimmune serum from the same animal. Binding took place at 37°C for 3-4 hours. The wells were washed as above and secondary antibody (goat anti-rabbit IgG conjugated to alkaline phosphatase) added at a concentration of 1 μg/ml, using the same conditions as the primary antibody. The wells are again washed, and alkaline phosphatase substrate (p-nitrophenyl phosphate; 0.6 mg/ml in diethanol-amine buffer pH 9.8) added. The yellow colour, indicating a positive reaction, was allowed to develop for 30 minutes and the reaction stopped with 3M NaOH. The colour is quantified at 405 nm. More detail of this method is given in Hill, 1984. The method confirmed that the antibodies all had a high titre and could be used at 1 μg/ml concentration.

Example 4

Isolation of Total RNA from Immature Cocoa Beans

The starting material for RNA which should contain a high proportion of mRNA specific for the storage proteins was immature cocoa beans, at about 130 days after pollination. Previous work had suggested that synthesis of storage proteins was approaching its height by this date (Biehl *et al*, 1982). The beans are roughly corrugated and pale pinkish-purple at this age.

The initial requirement of the total RNA preparation from cocoa beans was that it should be free from contaminants, as judged by the UV spectrum, particularly in the far UV, where a deep trough at 230 nm (260 nm : 230 nm ratio is approximately 2.0) is highly diagnostic of clean RNA, and is intact, as judged by agarose gel electrophoresis of heat-denatured samples, which should show clear rRNA bands. A prerequisite for obtaining intact RNA is scrupulous cleanliness and rigorous precautions against RNases, which are ubiquitous and extremely stable enzymes. Glassware is customarily baked at high temperatures, and solutions and apparatus treated with the RNase inhibitor diethyl pyrocarbonate (DEPC, 0.1%) before autoclaving.

The most routine method for extraction of plant (and animal) RNA is extraction of the proteins with phenol/chloroform in the presence of SDS to disrupt protein-nucleic acid complexes, and inhibit the RNases which are abundant in plant material. Following phenol extraction the RNA is pelletted on a caesium chloride gradient before or after ethanol precipitation. This method produced more or less intact RNA, but it was heavily contaminated with dark brown pigment, probably oxidised polyphenols and tannins, which always co-purified with the RNA. High levels of polyphenols are a major problem in *Theobroma* tissues.

A method was therefore adopted which avoided the use of phenol, and instead used the method of Hall *et al.* (1978) which involves breaking the tissue in hot SDS-borate buffer, digesting the proteins with proteinase K, and specifically precipitating the RNA with LiCl. This method gave high yields of reasonably clean, intact RNA. Contaminants continued to be a problem and the method was modified by introducing repeated LiCl precipitation steps, the precipitate being dissolved in water and clarified by microcentrifugation after each step. This resulted in RNA preparations with ideal spectra, which performed well in subsequent functional tests such as *in vitro* translation.

Preparation of mRNA From Total RNA

The mRNA fraction was separated from total RNA by affinity chromatography on a small (1 ml) oligo-dT column, the mRNA binding to the column by its poly A tail. The RNA (1-2 mg) was denatured by heating at 65°C and applied to the column in a high salt buffer. Poly A+ was eluted with low salt buffer, and collected by ethanol precipitation. The method is essentially that of Aviv and Leder (1972), modified by Maniatis *et al* (1982). From 1 mg of total RNA, approximately 10-20 μg polyA+ RNA was obtained (1-2%).

In vitro *Translation of mRNA*

The ability of mRNA to support *in vitro* translation is a good indication of its cleanliness and intactness. Only mRNAs with an intact polyA tail (3' end) will be selected by the oligo-dT column, and only mRNAs which also have an intact 5' end (translational start) will translate efficiently. *In vitro* translation was carried out using RNA-depleted wheat-germ lysate (Amersham International), the *de novo* protein synthesis being monitored by the incorporation of [^{35}S]-methionine (Roberts and Paterson, 1973). Initially the rate of *de novo* synthesis was measured by the incorporation of [^{35}S]-methionine into TCA-precipitable material trapped on glass fibre filters (GFC, Whatman). The actual products of translation were investigated by running on SDS-PAGE, soaking the gel in fluor, drying the gel and autoradiography. The mRNA preparations translated efficiently and the products covered a wide range of molecular weights, showing that intact mRNAs for even the largest proteins had been obtained. None of the major translation products corresponded in size to the 47kD or 31kD storage polypeptides identified in mature beans, and it was apparent that considerable processing of the nascent polypeptides must occur to give the mature forms.

Example 5

Identification of Precursor to the 47 kD and 31 kD Polypeptides by Immunoprecipitation

Because the 47 kD and 31 kD storage polypeptides were not apparent amongst the translation products of mRNA from developing cocoa beans, the technique of immunoprecipitation, with specific antibodies raised to the storage polypeptides, was used to identify the precursors from the translation mixture. This was done for two reasons: first to confirm that the appropriate mRNA was present before cloning, and second to gain information on the expected size of the encoding genes.

Immunoprecipitation was by the method of Cuming *et al*, 1986. [^{35} S]-labelled *in vitro* translation products were dissociated in SDS, and allowed to bind with specific antibody in PBS plus 1% BSA. The antibody-antigen mixture was then mixed with protein A-SEPHAROSE and incubated on ice to allow the IgG to bind to protein A. The slurry was poured into a disposable 1 ml syringe, and unbound proteins removed by washing with PBS +1% NONIDET P-40. The bound antibody was eluted with 1M acetic acid and the proteins precipitated with TCA. The antibody-antigen complex was dissociated in SDS, and subject to SDS-PAGE and fluorography, which reveals which labelled antigens have bound to the specific antibodies.

The results showed that the anti-47 kD and anti-31 kD antibodies both precipitated a 67 kD precursor. The precursor size corresponded to a major band on the *in vitro* translation products. The results with the 47 kD and 31 kD antibodies confirmed that the polypeptides are derived from a single precursor, or at least precursors of the same size. The large size of the precursor suggested that size-selection at mRNA or cDNA level may be necessary to obtain clones.

Example 6

cDNA Synthesis From the mRNA Preparations

cDNA synthesis was carried out using a kit from Amersham International. The first strand of the cDNA is synthesised by the enzyme reverse transcriptase, using the four nucleotide bases found in DNA (dATP, dTTP, dGTP, dCTP) and an oligo-dT primer. The second strand synthesis was by the method of Gubler and Hoffman (1983), whereby the RNA strand is nicked in many positions by RNase H, and the remaining fragments used to prime the replacement synthesis of a new DNA strand directed by the enzyme *E. coli* DNA polymerase I. Any 3' overhanging ends of DNA are filled in using the enzyme T4 polymerase. The whole process was monitored by adding a small proportion of [^{32}P]-dCTP into the initial nucleotide mixture, and measuring the percentage incorporation of label into DNA. Assuming that cold nucleotides are incorporated at the same rate, and that the four bases are incorporated equally, an estimate of the synthesis of cDNA can be obtained. From 1 μg of mRNA approximately 140 ng of cDNA was synthesised. The products were analysed on an alkaline 1.4% agarose gel as described in the Amersham methods. Globin cDNA, synthesised as a control with the kit, was run on the same gel, which was dried down and autoradiographed. The cocoa cDNA had a range of molecular weights, with a substantial amount larger than the 600 bp of the globin cDNA.

Example 7

Cloning of cDNA into a Plasmid Vector by Homopolymer Tailing

The method of cloning cDNA into a plasmid vector was to 3' tail the cDNA with dC residues using the enzyme terminal transferase (Boehringer Corporation Ltd), and anneal into a *Pst*I-cut and 5' tailed plasmid (Maniatis *et al*, 1982 Eschenfeldt *et al*, 1987). The optimum length for the dC tail is 12-20 residues. The tailing reaction (conditions as described by the manufacturers) was tested

with a 1.5 kb blunt-ended restriction fragment, taking samples at intervals, and monitoring the incorporation of a small amount of [^{32}P]-dCTP. A sample of cDNA (70 ng) was then tailed using the predetermined conditions.

A dG-tailed plasmid vector (3'-oligo(dG)-tailed pUC9) was purchased from Pharmacia. 15 ng vector was annealed with 0.5 - 5 ng of cDNA at 58°C for 2 hours in annealing buffer: 5mM Tris-HCl pH 7.6; 1mM EDTA, 75 mM NaCl in a total volume of 50 μl. The annealed mixture was transformed into *E. coli* RRI (Bethesda Research Laboratories), transformants being selected on L-agar + 100 μg/ml ampicillin. Approximately 200 transformants per ng of cDNA were obtained. Transformants were stored by growing in 100 μl L-broth in the wells of microtitre plates. adding 100 μl 80% glycerol, and storing at -20°C.

Some of the dC tailed cDNA was size selected by electrophoresing on a 0.8% agarose gel, cutting slits in the gel at positions corresponding to 0.5, 1.0 and 1.5 kb, inserting DE81 paper and continuing electrophoresis until the cDNA had run onto the DE81 paper. The DNA was then eluted from the paper with high salt buffer, according to the method of Dretzen *et al* (1981).

Example 8

Construction of Oligonucleotide Probes for the 47/31 kD Gene

The amino-acid sequence obtained from a cyanogen bromide peptide common to the 47 kD and 31 kD polypeptides is as follows:

Met-Phe-Glu-Ala-Asn-Pro-Asn-Thr-Phe

and the least redundant probe of 17 residues (a mixture of 32) is shown below:

```
          Met-Phe-Glu-Ala-Asn-Pro
5'    ATG TTT GAA GCT AAT CC    3'
            C   G   C   C
                    A
                    G
```

The actual probe was made anti-sense so that it could also be used to probe mRNA. Probe synthesis was carried out using an Applied Biosystems apparatus.

Example 9

Use of Oligonucleotides to Probe cDNA Library

The oligonucleotide probes were 5' end-labelled with gamma-[^{32}P] dATP and the enzyme polynucleotide kinase (Amersham International). The method was essentially that of Woods (1982, 1984), except that a smaller amount of isotope (15 μCi) was used to label about 40 ng probe, in 10 mM $MgCl_2$, 100 mM Tris-HCl, pH 7.6; 20 mM 2-mercaptoethanol.

The cDNA library was grown on GeneScreen (New England Nuclear) nylon membranes placed on the surface of L-agar + 100 μg/ml ampicillin plates. (The word GeneScreen is a trade mark.) Colonies were transferred from microtitre plates to the membranes using a 6 x 8 multi-pronged device, designed to fit into the wells of half the microtitre plate. Colonies were grown overnight at 37°C, lysed in sodium hydroxide and bound to membranes as described by Woods (1982, 1984). After drying the membranes were washed extensively in 3 x SSC/0.1% SDS at 65°C, and hybridised to the labelled probe, using a HYBAID apparatus from Hybaid Ltd, PO Box 82, Twickenham, Middlesex. (The word HYBAID is a trade mark.) Conditions for hybridisation were as described by Mason & Williams (1985), a T_d being calculated for each oligonucleotide according to the formula:

T_d = 4°C per GC base pair + 2°C per AT base pair.
At mixed positions the lowest value is taken.

Hybridisation was carried out at T_d -5°C. Washing was in 6 x SSC, 0.1% SDS initially at room temperature in the HYBAID apparatus, then at the hybridisation temperature (T_d -5°C) for some hours, and finally at T_d for exactly 2 minutes. Membranes were autoradiographed onto FUJI X-ray film, with intensifying screens at - 70°C. (The word FUJI is a trade mark.) After 24 - 48 hours positive colonies stood out as intense spots against a low background.

Example 10

Analysis of Positive Clones for the 47 kD/31 kD Polypeptide

Only one positive clone, pMS600, was obtained. This released two *Pst*I fragments on digestion, of total length 1.3 kb, insufficient to encode the 67 kD precursor. The total insert was removed from the vector on a *Hin*dIII-*Eco*RI fragment, nick-translated and used to probe the cDNA library, picking up a further two positive clones, pMS700 and pMS800. Restriction mapping of all three inserts suggested an overlapping map covering nearly 2.0 kb, sufficient to encode the 67 kD precursor (see Figure 1).

Example 11

Sequencing the Cloned Inserts

The sequencing strategy was to clone the inserts, and where appropriate subclones thereof, into the multiple cloning site of the plasmids pTZ18R/pTZ19R (Pharmacia). These plasmids are based on the better-known vectors pUC18/19 (Norrander *et al.* 1983), but contain a single-stranded origin of replication from the filamentous phage f1. When superinfected with phages in the same group, the plasmid is induced to undergo single-stranded

replication, and the single-strands are packaged as phages extruded into the medium. DNA can be prepared from these 'phages' using established methods for M13 phages (Miller, 1987), and used for sequencing by the method of Sanger (1977) using the reverse sequencing primer. The superinfecting phage used is a derivative of M13 termed M13K07, which replicates poorly and so does not compete well with the plasmid, and contains a selectable kanamycin-resistance marker. Detailed methods for preparing single-strands from the pTZ plasmids and helper phages are supplied by Pharmacia. DNA sequence was compiled and analysed using the Staden package of programs (Staden, 1986), on a PRIME 9955 computer. (The word PRIME is a trade mark.)

Example 12

Features of the 47 kD/31 kD cDNA and Deduced Amino-acid Sequence of the 67 kD Precursor

DNA sequencing of the three positive clones, pMS600, pMS700, pMS800, confirmed the overlap presumed in Figure 1. No sequence differences were found in the overlapping regions (about 300 bp altogether), suggesting that the three cDNAs were derived from the same gene. The sequence of the combined cDNAs comprising 1818 bases is shown in Figure 2. The first ATG codon is found at position 14, and is followed by an open reading frame of 566 codons. There is a 104-base 3' untranslated region containing a polyadenylation signal at position 1764. The oligonucleotide probe sequence is found at position 569.

The open reading frame translates to give a polypeptide of 566 amino-acids (Figure 2), and a molecular weight of 65612, which is reasonably close to the 67 kD measured on SDS-PAGE gels. The N-terminal residues are clearly hydrophobic and look like a characteristic signal sequence. Applying the rules of Von Heije (1983), which predict cleavage sites for signal sequences, suggests a cleavage point between amino-acids 20 and 21 (see Figure 3). The region following this is highly hydrophobic and contains four Cys-X-X-X-Cys motifs.

The N-terminus of the mature protein has been roughly identified as the glutamate (E) residue at 135 (Figure 3), on the basis of some tentative N-terminal sequence (EEPGSQFANPAYHF). This N-terminus would give a mature protein of 49068 kD, in rough agreement with that observed. There appears to be no glycosylation sites (Asn-X-Ser/Thr) in the mature protein of the sequence.

Homologies Between the 67 kD Precursor and Other Known Proteins

Searches through the PIR database, and through the literature, revealed close homologies between the 67 kD polypeptide and a class of seed storage proteins termed vicilins, one of two major classes of globulins found in seed (Borroto and Dure, 1987). Alignments between the 67 kD polypeptide and vicilins from cotton (*Gossypium hirsutum*, Ghi), soybean (*Glycine max*, Gma), pea (*Pisum sativum*, Psa-c is convicilin, Psa-v is vicilin) and bean (*Phaseolus vulgaris*, Pvu) are shown in Figure 4 (Bown *et al*, 1988; Chlan *et al*, 1986; Doyle *et al*, 1986; Lycett *et al*, 1983). Identical residues are boxed.

All the vicilins have a mature molecular weight of around 47 kD, with the exception of soybean conglycinin alpha and alpha1 subunits, which are 72 kD and 76 kD respectively, and pea convicilin with a mature molecular weight of 64kD. The pea and bean subunits (2 sub classes each) are synthesised as small precursors, around 50 kD. The most striking homology with the 67 kD is the cotton vicilin (Chlan *et al*, 1986). Cotton is also the most closely related to cocoa: both are members of the order Malvales. Interestingly cotton also has a large precursor, of 69 kD, comprising a short signal sequence, a large hydrophilic domain containing six Cys-X-X-X-Cys motifs, and a mature domain. It may therefore be possible to synthesise the corresponding cotton protein, by techniques analogous to those disclosed in this application and to use the cotton protein, or fragments of it, in the preparation of flavour components analogous to cocoa flavour components.

Example 13

Expression of the 67 kD Polypeptide in E. coli

Before the 67 kD coding region could be inserted into a expression vector the overlapping fragments from the three separate positive clones had to be spliced into a continuous DNA segment. The method of splicing is illustrated in Figure 6: a *Hin*dIII-*Bgl*II fragment from pMS600, a *Bgl*II-*Eco*RI fragment from pMS700 and an *Eco*RI-*Sal*I fragment from pMS800 were ligated into pTZ19R cut with *Hin*dIII and *Sal*I. The resulting plasmid, containing the entire 67 kD cDNA, was termed pMS900.

An *Nco*I site was introduced at the ATG start codon, using the mutagenic primer:

5' TAG CAA CCA TGG TGA TCA 3'.

In vitro mutagenesis was carried out using a kit marketed by Amersham International, which used the method of Eckstein and co-workers (Taylor *et al*, 1985). After annealing the mutagenic primer to single-stranded DNA the second strand synthesis incorporates alpha-thio-dCTP in place of dCTP. After extension and ligation to form closed circles, the plasmid is digested with *Nci*I, an enzyme which cannot nick DNA containing thio-dC. Thus only the original strand is nicked, and subsequently digested with exonuclease III. The original strand is then resynthesised, primed by the remaining DNA fragments and complementing the mutated position in the original strand. Plasmids are then transformed into *E. coli* and checked by plasmid mini preparations.

The 67 kD cDNA was then cloned into the *E. coli* expression plasmid, pJLA502 (Figure 5), on an *Nco*I - *Sal*I fragment (pMS902).

pJLA502 (Schauder *et al*, 1987) is marketed by Medac GmbH, Postfach 303629, D-7000, Hamburg 36 and contains the strong lambda promoters, P_L and P_R, and the leader sequence and ribosome binding site of the very efficiently translated *E. coli* gene, *atp*E. It also contains a temperature-sensitive cI repressor, and so expression is repressed at 30°C and activated at 42°C. The vector has an *Nco*I site (containing an ATG codon: CCATGG) correctly placed with respect to the ribosome binding site, and foreign coding sequences must be spliced in at this point.

The expression vector was transformed into *E. coli* UT580. The transformed strain was grown in L-broth + ampicillin (100 μg/ml) at 30°C until log phase (OD_{610} = 0.5) and the temperature was then shifted to 42°C and samples taken at intervals. Samples were dissociated by boiling in SDS loading buffer, and run on SDS-PAGE gels. The proteins were electroblotted onto nitrocellulose membranes (Towbin *et al*, 1979) and Western blotting carried out using the anti-21 kD antibody prepared in Example 3 above (at 2 μg/ml) and as a secondary antibody, goat anti-rabbit-IgG conjugated to alkaline phosphatase (Scott *et al*, 1988).

A specific band at 67 kD was produced by pMS902, showing that a functional gene was present.

Example 14

Expression of the 67 kD Polypeptide in Yeast

Two yeast expression vectors were used, both based on a yeast-*E.coli* shuttle vector containing yeast and *E.coli* origins of replication, and suitable selectable markers (ampicillin-resistance for *E.coli* and leucine auxotrophy for yeast). Both vectors contain the yeast pyruvate kinase (PK) promoter and leader sequence and have a *Hin*dIII cloning site downstream of the promoter. One vector, A (YVA), is designed for internal expression, and the other, B (YVB), for secreted expression, having a portion of the signal sequence of the yeast

mating alpha-factor downstream of the promoter, with a *Hin*dIII site within it to create fusion proteins with incoming coding sequences. The vectors are illustrated in Figure 7.

To use the vectors effectively it is desirable to introduce the foreign coding region such that for vector A, the region from the *Hin*dIII cloning site to the ATG start is the same as the yeast PK gene, and for vector B, the remainder of the alpha-factor signal, including the lysine at the cleavage point. In practice this situation was achieved by synthesising two sets of *Hin*dIII - *Nco*I linkers to breach the gap between the *Hin*dIII cloning site in the vector and the *Nco*I at the ATG start of the coding sequence. This is illustrated in Figure 8.

In order to use the yeast vector B, the hydrophobic signal sequence must first be deleted from the 67 kD cDNA. Although direct evidence of the location of the natural cleavage site was lacking, the algorithm of Von Heije predicts a site between amino-acids 20 (alanine) and 21 (leucine). However it was decided to remove amino-acids 2-19 by deletion, so that the useful *Nco*I site at the translation start would be maintained.

For ease of construction of the yeast vectors, the strategy was to first clone the *Hin*dIII - *Nco*I linkers into the appropriate pTZ plasmids, and then to clone the linkers plus coding region into the yeast vectors on *Hin*dIII - *Bam*HI fragments. However the coding region contains an internal *Bam*HI which must be removed by in vitro mutagenesis, giving a new plasmid pMS903. The signal sequence was deleted from pMS903 using the mutagenic primer

5' AGCATAGCAACCATGGTTGCTTTGTTCT 3'

to give pMS904. The appropriate *Hin*dIII - *Nco*I linkers were then cloned into pMS903 and pMS904 to give pMS907 and pMS905 respectively, and the *Hin*dIII - *Bam*HI fragments (linkers + coding region) subcloned from these

intermediate plasmids into YVA and YVB respectively to give the yeast expression plasmids pMS908 and pMS906. A diagrammatic scheme for these and other constructs is given in Figure 9.

Because the mature cocoa protein appears to lack the N-terminal hydrophilic domain, as described in Example 12, expression vectors have also been designed to express the mature protein directly. Yeast is unlikely to have the same processing enzymes as cocoa and optimum expression may be obtained for a protein as close as possible to that found naturally in cocoa. Hence the DNA encoding the hydrophilic domain (amino acids 20-134) was deleted from the intermediate plasmids pMS907 and pMS905 to give plasmids pMS911 and pMS909 respectively, and the *Hin*dIII - *Bam*HI fragments for these were cloned into YVA and YVB to give the expression plasmids pMS912 and pMS910 (Figure 9).

A further modification was introduced by constructing expression in which the plant terminator had been removed and replaced with the yeast ADH terminator (present in YVA and YVB). The plant signal was removed by cutting the intermediate plasmids pMS907 and pMS905 at the *Pvu*II site immediately downstream of the coding region, at position 1716 in Figure 2. *Hin*dIII linkers were added and the entire coding region cloned into the yeast expression vectors on *Hin*dIII - *Hin*dIII fragments giving expression plasmids pMS914 (YVA) and pMS916 (YVB) (Figure 9). A summary of the constructs made is given in Figure 10.

The yeast expression plasmids were transferred into yeast spheroplasts using the method of Johnston (1988). The transformation host was the LEU^- strain AH22, and transformants were selected on leucine-minus minimal medium. LEU^+ transformants were streaked to single colonies, which were grown in 50 ml YEPD medium (Johnston, 1988) at 28 | C for testing the extent and distribution of foreign protein. Cells were harvested from cultures in preweighed tubes in a bench-top centrifuge, and washed in 10 ml lysis buffer (200 mM Tris, pH 8.1: 10% glycerol). The cell medium was reserved and

concentrated 10-25 x in an AMICON mini concentrator. (The word AMICON is a trade mark.) The washed cells were weighed and resuspended in lysis buffer plus protease inhibitors (1 mM phenyl methyl sulphonyl fluoride (PMSF); 1 μg/ml aprotinin; 0.5 μg/ml leupeptin) at a concentration of 1 g/ml. 1 volume acid-washed glass-beads was added and the cells broken by vortexing for 8 minutes in total, in 1 minute bursts, with 1 minute intervals on ice. After checking under the microscope for cell breakage, the mixture was centrifuged at 7000 rpm for 3 minutes to pellet the glass beads. The supernatant was removed to a pre-chilled centrifuge tube, and centrifuged for 1 hour at 20,000 rpm. (Small samples can be centrifuged in a microcentrifuge in the cold.) The supernatant constitutes the soluble fraction. The pellet was resuspended in 1 ml lysis buffer plus 10% SDS and 1% mercaptoethanol and heated at 90°C for 10 minutes. After centrifuging for 15 minutes in a microcentrifuge the supernatant constitutes the particulate fraction.

Samples of each fraction and the concentrated medium were examined by Western blotting. Considering first the plasmids designed for internal expression in YVA, pMS908 produced immunoreactive proteins at 67 kD and 16 kD within the cells only. There was no evidence of the 67 kD protein being secreted under the influence of its own signal sequence. The smaller protein is presumed to be a degradation product. A similar result, but with improved expression, was obtained with pMS914, in which the plant terminator is replaced by a yeast terminator. However in pMS912, in which the coding region for the hydrophilic domain has been deleted, no synthesis of immunoreactive protein occurred.

For industrial production of heterologous proteins in yeast a secreted mode is preferable because yeast cells are very difficult to disrupt, and downstream processing from total cell protein is not easy. The results from the vectors constructed for secreted expressed were rather complicated. From the simplest construct, pMS906, in which the yeast α-factor signal sequence replaces the plant protein's own signal, immunoreactive proteins of approximately 47 kD, 28 kD and 18-20 kD were obtained and secreted into the medium. At first sight

this is surprising because the coding region introduced should synthesise a 67 kD protein. However the most likely explanation is that the yeast's KEX2 protease, that recognises and cleaves the α-factor signal at a Lys-Arg site is also cleaving the 67 KD protein at Lys-Arg dipeptides at positions 148 and 313 in the amino-acid sequence. The calculated protein fragment sizes resulting from cleavage at these positions are 47179 Daltons, 28344 Daltons and 18835 Daltons, very close to the observed sizes.

When the plant terminator is replaced with a yeast terminator in pMS916 no expression is obtained in either cells or medium; it is possible that a mutation has been inadvertantly introduced. From the construct pMS910, in which the hydrophilic domain has been deleted the main antigenic products were 28 KD and 18-20 kD, again secreted into the medium. It is presumed that the *de novo* 47 kD product is immediately cleaved at the KEX2 site at position 313.

In summary, four of the six expression vectors constructed direct the synthesis of proteins cross-reacting with anti-47 kD antibodies. Two of the constructs secrete proteins into the medium.

Example 15

Construction of Vectors Designed to Express the 67 kD Protein in Hansenula polymorpha

The methylotropic yeast *Hansenula polymorpha* offers a number of advantages over *Saccharamyces cerevisiae* as a host for the expression of heterologous proteins (EP-A-0173378 and Sudbery *et al*, 1988). The yeast will grow on methanol as sole carbon source, and under these conditions the enzyme methanol oxidase (MOX) can represent up to 40% of the total cell protein. Thus the MOX promoter is a very powerful one that can be used in a vector to drive the synthesis of heterologous proteins, and it is effective even as a single copy. This gives the potential to use stable integrated vectors. *Hansenula* can also grow on rich carbon sources such as glucose, in which case the MOX

promoter is completely repressed. This means that cells containing the heterologous gene can be grown to a high density on glucose, and induced to produce the foreign protein by allowing the glucose to run out and adding methanol.

A plasmid, pHGL1, containing the MOX promoter and terminator, and a cassette containing the yeast α-factor secretory signal sequence, were prepared. The 67 kD coding region was cloned into pHGL1 on a *Bam*HI - *Bam*HI fragment, replacing the *Bgl*II fragment which contains the 3' end of the MOX coding region. The whole promoter - gene - terminator region can then be transferred to YEp13 on a *Bam*HI - *Bam*HI fragment to give the expression plasmid pMS922. The details of the construction are illustrated in Figure 11. An analogous expression plasmid, pMS925, has been constructed with the yeast α-factor spliced onto the 67 kD coding region, replacing the natural plant signal. The *Bam*HI - *Hin*dIII cassette containing the α-factor was ligated to the *Hin*dIII - *Bam*HI fragment used to introduce the 67 kD coding region into YVB. The α-factor plus coding region was then cloned with pHGL1 on a *Bam*HI - *Bam*HI fragment, and transferred into YEP13 as before. Details are shown in Figure 12.

Both constructs have been transformed into *Hansenula* and grown under inducing conditions with 0.5% or 1% methanol. Both constructs directed the production of immunoreactive protein within the cells, and pMS925 secreted the protein into the medium under the influence of the α-factor signal sequence.

E. coli Strains

RR1	$F^-v_B^-M_B$ *ara*-14 *pro*A2 *leu*B6 *lac*Y1 *gal*K2 *vps*L20 (str^r) *xyl*-5 *mtl*-1 *sup*E44
CAG629	lac_{am} tvp_{am} pho_{am} $htpR_{am}$ *mal* *rps*L *lon* $sup C_{ts}$

UT580 (*lac-pro*) *sup*E *thi* *hsd*D5 / F'*tra* D36 *pro*A$^+$B$^+$ *lac*I^q *lac*Z M15

References

Aviv, H., and Leder, P. *Proc. Natl. Acad. Sci. USA* **69**, 1408-1412 (1972). Purification of biologically active globin mRNA by chromatography on oligo dT cellulose

Biehl, B., Wewetzer, C., and Passern, D. *J. Sci. Food Agric.* **33**, 1291-1304 (1982). Vacuolar (Storage) Proteins of Cocoa Seeds and their Degradation during Germination and Fermentation.

Borroto, K., and Dure, L. *Plant Molecular Biology* **8**, 113-131 (1987). The globulin seed storage proteins of flowering plants are derived from two ancestral genes.

Bown.D., Ellis. T.H.N., and Gatehouse, J.A. *Biochem. J.* **251**, 717-726 (1988). The sequence of a gene encoding convicilin from pea (*Pisum sativum*) shows that convicilin differs from vicilin by an insertion near the N-terminus.

Catty, D. and Raykundalia, C. Production and Quality control of Polyclonal Antibodies in: "Antibodies: A Practical Approach" Vol I, IRL Press (1988)

Chlan, C.A., Pyle, J.B., Legocki, A.B., and Dure, L. *Plant Molecular Biology* 7, 475-489 (1986). Developmental biochemistry of cotton seed embryogenesis and germination XVIII. cDNA and amino acid sequences of members of the storage protein families.

Cuming, A.C., Williams. R.S., and Cullimore, J.V. in "Immunology in Plant Science", Ed. Wang, T.L., Cambridge University Press, 1986. The use of Antibodies in Molecular Biology.

Doyle, J.J., Schuler, M.A., Godette, W.D., Zencer, V., Beachy, R.N., and Slightom, J.L. *J. Biol. Chem.* **261**, 9228-9238 (1986). The glycosylated seed storage proteins of Glycine max and Phaseolus vulgaris : Structural homologies of genes and proteins.

Dretzen, G., Bellard, M., Sassone-Corsi, P., and Chambon, P. *Analytical Biochemistry* **112**, 295-298 (1981). A reliable method for the recovery of DNA fragments from agarose and acrylamide gels.

Eschenfeldt, W.H., Puskas, R.S., and Berger, S.L. *Methods in Enzymology* **152**, 337-342 (1987). Homopolymeric Tailing.

Fritz *et al* (*J. Food Sci.* **50** 946-950 (1985))

Gubler, U., and Hoffman, B.J. *Gene* **25**, 263 (1983). A simple and very efficient method for generating cDNA libraries.

Hall, T.C., Ma, Y., Buchbinder, B.U., Pyrne J.W., Sun, S.M., and Bliss, F.A. *Proc. Natl. Acad. Sci. USA* **75**, 3196-3200 (1978). Messenger RNA for G1 protein of French bean seeds : cell-free translation and product characterisation.

Hill, S.A. "Methods in Plant Virology", Blackwell 1984.

Johnston, J.R. in "Yeast : A practical approach". Eds Campbell, I., and Duffus, J.H. IRL Press. 1988. Yeast Genetics. Molecular Aspects.

Kreil, G. *Annual Rev. Biochem.* **50**, 317-348 (1981). Transfer of proteins across membranes.

Laemmli, U.K. *Nature* **227**, 680 (1970). Cleavage of structural proteins during the assembly of the head of bacteriophage T4.

Lipman, D.J., and Pearson, W.R. *Science* **227**, 1435-1441 (1985). Rapid protein sequence similarity searches.

Lycett, G.W., Delauney, A.J., Gatehouse, J.A., Gilroy, J., Croy, R.R.D., and Boulter, D. *Nucleic Acids Res.* **11**, 2367-2380 (1983).

Maniatis, T., Fritsch, E.F., and Sambrook, J. "Molecular Cloning: A Laboratory Manual", Cold Spring Harbour Laboratory, 1982.

Mason, P.J., and Williams, J.G. in "Nucleic Acid Hybridisation : A Practical Approach". Ed. Hames, B.D., and Higgins, S.J. IRL Press 1985. Hybridisation in the Analysis of Recombinant DNA.

Meinkoth, J., and Wahl, G.M. *Analytical Biochemistry* **138**, 267 (1984). Methods of Southern blotting and DNA probing.

Miller, H. *Methods in Enzymology* **152**, 145-170 (1987). Practical Aspects of Preparing Phage and Plasmid DNA : Growth, Maintenance and Storage of Bacteria and Bacteriophage.

Norrander, J., Kempe, T., and Messing, J. *Gene* **26**, 101 (1983). Construction of improved M13 vectors using oligodeoxynucleotide-directed mutagensis.

Proudfoot, N.J., and Brownlee, G.G. *Nature* **263**, 211-214 (1976). 3' Non-coding region sequences in enkayotic messenger RNA.

Roberts, B.E., and Paterson, B.M. *Proc. Natl. Acad. Sci. USA* **70**, 2330 (1973). Efficient translation of tobacco mosaic virus RNA and rabbit globin 9S RNA in a cell-free system from commercial wheat germ.

Sanger, F., Nicklen, S., and Coulson, A.R. *Proc. Natl. Acad. Sci. USA* **74**, 5463-5467 (1977). DNA sequencing with chain-terminating inhibitors.

Schauder, B., Blocker, H., Frank, R., and McCarthy, J.E.G. *Gene* **52**, 279-283 (1987). Inducible expression vectors incorporating the *E. coli apt*E translation initiation region.

Scott, R., Draper, J., Jefferson, R., Dury, G., and Jacob, L. in "Plant Genetic Transformation and Gene Expression: A Laboratory Manual". Eds. Draper, J., Scott, R., Armitage, P., Walden, R. Blackwell 1988. Analysis of gene organisation and expression in plants.

Staden, R. *Nucleic Acids Res.* **14**, 217-231 (1986). The current status and portability of our sequence handling software.

Sudbery, P.E., Gleeson, M.A., Veale, R.A., Lederboer, A.M., and Zoetmulder, M.C.M. *Biocem. Soc. Trans* 16 1081-103 (1988). Hansenula polymorpha as a novel yeast system for the expression of heterologous genes.

Taylor, J.W., Ott, J., and Eckstein, F. *Nucleic Acids Res.* **13**, 8765-8785 (1985). The rapid generation of oligonucleotide-directed mutations at high frequency using phosphorothioate-modified DNA.

Towbin, H., Staehelin, T., and Gordon, J. *Proc. Natl. Acad. Sci. USA* **76**, 4350-4534 (1979). Electrophoretic transfer of proteins from polyacrylamide gels to nitrocellulose sheets : procedure and some applications.

Von Heije, G. *Eur. J. Biochem* **133**, 17-21 (1983). Patterns of Amino-acids near Signal-Sequence Cleavage Sites.

Wahl, G.M., and Berger, S.L. *Methods in Enzymology* **152**, 415-423 (1987). Screening Colonies or Plaques with Radioactive Nucleic Acid Probes.

Woods, D.E. *Focus (Bethesda Research Labs)* **6**, 3 (1984). Oligonucleotide Screening of cDNA Libraries.

Woods, D.E., Markham, A.F., Ricker, A.T., Goldberger, G., and Colten, H.R. *Proc. Natl. Acad. Sci. USA* **79**, 5561 (1982).

CLAIMS

1. A 67kD protein of *Theobroma cacao*, or a fragment thereof.

2. A 47kD protein of *Th. cacao*, or a fragment thereof.

3. A 31kD protein of *Th. cacao*, or a fragment thereof.

4. A protein as claimed in claim 1, 2 or 3, having at least part of the sequence shown in Figure 2.

5. A fragment as claimed in any one of claims 1 to 4, which comprises at least four amino acids.

6. A protein or fragment as claimed in any one of claims 1 to 6, which is recombinant.

7. Recombinant or isolated nucleic acid coding for a protein or fragment as claimed in any one of claims 1 to 5.

8. Nucleic acid as claimed in claim 7 which is DNA.

9. Nucleic acid as claimed in claim 8, having at least part of the sequence shown in Figure 2.

10. Nucleic acid as claimed in claim 7, 8 or 9, which is in the form of a vector.

11. Nucleic acid as claimed in claim 10, wherein the vector is an expression vector and the protein- or fragment-coding sequence is operably linked to a promoter.

12. Nucleic acid as claimed in claim 11, wherein the expression vector is a yeast expression vector and the promoter is a yeast pyruvate kinase (PK) promoter.

13. Nucleic acid as claimed in claim 11, wherein the expression vector is a bacterial expression vector and the promoter is a strong lambda promoter.

14. Nucleic acid as claimed in claim 11, 12 or 13, comprising a signal sequence.

15. A host cell comprising nucleic acid as claimed in any one of claims 10 to 14.

16. A host cell as claimed in claim 15 which is *Saccharomyces cerevisiae*.

17. A host cell as claimed in claim 15 which is *E. coli*.

18. A process for the preparation of a protein or fragment as claimed in any one of claims 1 to 5, the process comprising coupling successive amino acids by peptide bond formation.

19. A process for the preparation of a protein or fragment as claimed in any one of claims 1 to 5, the process comprising culturing a host cell as claimed in claim 15, 16 or 17.

20. A process for the preparation of nucleic acid as claimed in any one of claims 7 to 14, the process comprising coupling together successive nucleotides and/or ligating oligo- or poly-nucleotides.

FIG. 1

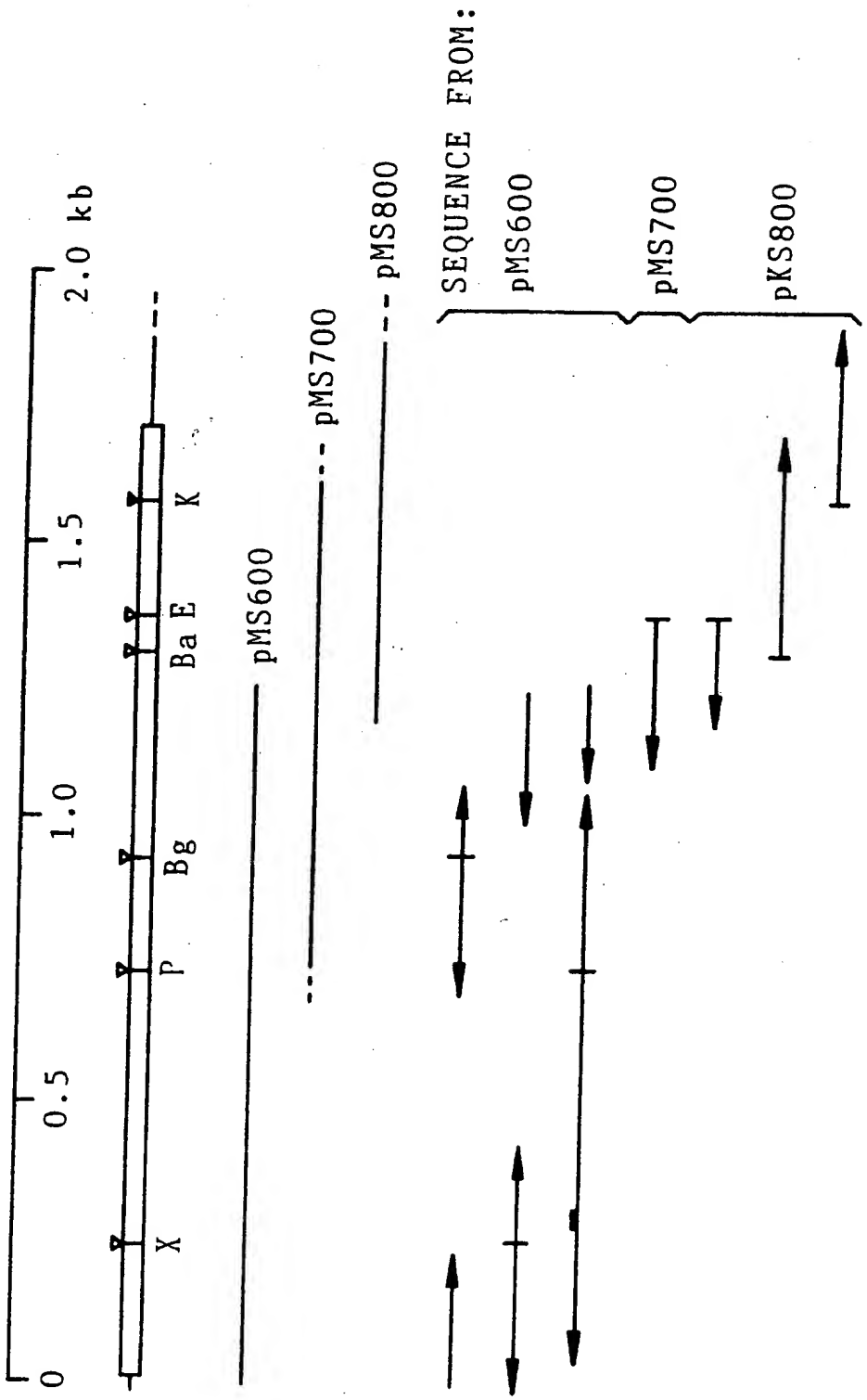


FIG. 2A

```
                  M  V  I  S  K  S  P  F  I  V  L  I  F  S  L  L
AAGCATAGCAAATATGGTGATCAGTAAGTCTCCTTTCATAGTTTTGATCTTCTCTCTTCT
        10         20         30         40         50         60

   L  S  F  A  L  L  C  S  G  V  S  A  Y  G  R  K  Q  Y  E  R
CCTTTCTTTTGCGTTGCTTTGTTCTGGTGTCAGCGCCTATGGCAGAAAACAATATGAGCG
        70         80         90        100        110        120

   D  P  R  Q  Q  Y  E  Q  C  Q  R  R  C  E  S  E  A  T  E  E
TGATCCTCGACAGCAATACGAGCAATGCCAGAGGCGATGCGAGTCGGAAGCGACTGAAGA
       130        140        150        160        170        180

   R  E  Q  E  Q  C  E  Q  R  C  E  R  E  Y  K  E  Q  Q  R  Q
AAGGGAGCAAGAGCAGTGTGAACAACGCTGTGAAAGGGAGTACAAGGAGCAGCAGAGACA
       190        200        210        220        230        240

   Q  E  E  E  L  Q  R  Q  Y  Q  Q  C  Q  G  R  C  Q  E  Q  Q
GCAAGAAGAAGAGCTTCAAAGGCAATACCAGCAATGTCAAGGGCGTTGTCAAGAGCAACA
       250        260        270        280        290        300

   Q  G  Q  R  E  Q  Q  Q  C  Q  R  K  C  W  E  Q  Y  K  E  Q
ACAGGGGCAGAGAGAGCAGCAGCAGTGCCAGAGAAAATGCTGGGAGCAATATAAGGAACA
       310        320        330        340        350        360

   E  R  G  E  H  E  N  Y  H  N  H  K  K  N  R  S  E  E  E  E
AGAGAGAGGCGAGCACGAGAATTACCATAATCACAAAAAAAATAGGAGCGAAGAAGAAGA
       370        380        390        400        410        420

   G  Q  Q  R  N  N  P  Y  Y  F  P  K  R  R  S  F  Q  T  R  F
AGGGCAACAAAGAAACAATCCTTACTATTTTCCTAAAAGAAGATCATTCCAAACTCGATT
       430        440        450        460        470        480
```

FIG. 2B

```
  R  D  E  E  G  N  F  K  I  L  Q  R  F  A  E  N  S  P  P  L
CAGGGATGAAGAGGGCAACTTCAAGATCCTCCAGAGGTTTGCTGAGAACTCTCCTCCACT
       490        500        510        520        530        540

  K  G  I  N  D  Y  R  L  A  M  F  E  A  N  P  N  T  F  I  L
CAAGGGCATCAACGATTACCGCTTGGCCATGTTCGAAGCAAATCCCAACACTTTTATTCT
       550        560        570        580        590        600

  P  H  H  C  D  A  E  A  I  Y  F  V  T  N  G  K  G  T  I  T
TCCGCACCACTGTGATGCTGAGGCAATTTACTTCGTGACAAACGGAAAGGGGACAATTAC
       610        620        630        640        650        660

  F  V  T  H  E  N  K  E  S  Y  N  V  Q  R  G  T  V  V  S  V
GTTTGTGACTCATGAAAACAAAGAGTCCTATAATGTACAGCGTGGAACAGTAGTCAGCGT
       670        680        690        700        710        720

  P  A  G  S  T  V  Y  V  V  S  Q  D  N  Q  E  K  L  T  I  A
TCCTGCAGGAAGCACTGTTTACGTGGTTAGCCAAGACAACCAAGAGAAGCTAACCATAGC
       730        740        750        760        770        780

  V  L  A  L  P  V  N  S  P  G  K  Y  E  L  F  F  P  A  G  N
TGTGCTCGCCCTGCCTGTTAATTCTCCTGGCAAATATGAGTTATTCTTCCCCGCTGGAAA
       790        800        810        820        830        840

  N  K  P  E  S  Y  Y  G  A  F  S  Y  E  V  L  E  T  V  F  N
TAATAAACCTGAATCATATTACGGAGCCTTCAGCTATGAAGTTCTTGAGACCGTCTTCAA
       850        860        870        880        890        900

  T  Q  R  E  K  L  E  E  I  L  E  E  Q  R  G  Q  K  R  Q  Q
TACACAAAGAGAGAAGCTGGAGGAGATCTTGGAGGAACAGAGAGGGCAGAAGAGGCAGCA
       910        920        930        940        950        960
```

FIG. 2C

```
  G  Q  Q  G  M  F  R  K  A  K  P  E  Q  I  R  A  I  S  Q  Q
GGGGCAGCAGGGTATGTTCCGGAAAGCCAAACCAGAGCAGATAAGAGCAATAAGCCAACA
      970        980        990       1000       1010       1020

  A  T  S  P  R  H  R  G  G  E  R  L  A  I  N  L  L  S  Q  S
AGCTACTTCTCCAAGGCACAGAGGCGGGGAGAGACTTGCCATCAATCTATTGAGCCAATC
     1030       1040       1050       1060       1070       1080

  P  V  Y  S  N  Q  N  G  R  F  F  E  A  C  P  E  D  F  S  Q
GCCTGTCTACTCCAACCAAAACGGACGCTTCTTTGAGGCTTGTCCTGAGGACTTCAGTCA
     1090       1100       1110       1120       1130       1140

  F  Q  N  M  D  V  A  V  S  A  F  K  L  N  Q  G  A  I  F  V
ATTTCAGAACATGGATGTCGCTGTTTCAGCCTTCAAACTGAATCAGGGAGCCATATTTGT
     1150       1160       1170       1180       1190       1200

  P  H  Y  N  S  K  A  T  F  V  V  F  V  T  D  G  Y  G  Y  A
GCCACACTACAATTCTAAGGCTACATTCGTGGTGTTTGTCACGGACGGATATGGGTACGC
     1210       1220       1230       1240       1250       1260

  Q  M  A  C  P  H  L  S  R  Q  S  Q  G  S  Q  S  G  R  Q  D
TCAAATGGCTTGCCCGCATCTCTCCAGACAGAGCCAGGGATCCCAAAGTGGAAGGCAAGA
     1270       1280       1290       1300       1310       1320

  R  R  E  Q  E  E  E  S  E  E  E  T  F  G  E  F  Q  Q  V  K
CAGAAGAGAACAAGAAGAAGAGTCAGAAGAGGAGACATTTGGAGAATTCCAGCAGGTCAA
     1330       1340       1350       1360       1370       1380

  A  P  L  S  P  G  D  V  F  V  A  P  A  G  H  A  V  T  F  F
AGCCCCATTGTCACCTGGTGACGTCTTTGTAGCCCCGGCAGGCCATGCAGTTACATTCTT
     1390       1400       1410       1420       1430       1440
```

FIG. 2D

```
  A  S  K  D  Q  P  L  N  A  V  A  F  G  L  N  A  Q  N  N  Q
TGCATCCAAAGACCAGCCCCTGAATGCAGTTGCGTTTGGACTCAACGCCCAGAACAACCA
      1450       1460       1470       1480       1490       1500

  R  I  F  L  A  G  K  K  N  L  V  R  Q  M  D  S  E  A  K  E
GAGAATTTTCCTTGCAGGGAAAAAGAACTTGGTCAGACAAATGGATAGCGAGGCAAAGGA
      1510       1520       1530       1540       1550       1560

  L  S  F  G  V  P  S  K  L  V  D  N  I  F  N  N  P  D  E  S
GTTATCATTTGGGGTACCATCGAAATTGGTAGATAATATATTCAACAACCCGGATGAGTC
      1570       1580       1590       1600       1610       1620

  Y  F  M  S  F  S  Q  Q  R  Q  R  R  D  E  R  R  G  N  P  L
GTATTTCATGTCTTTCTCTCAACAGAGGCAGCGTCGAGATGAAAGGAGGGGCAATCCCTT
      1630       1640       1650       1660       1670       1680

  A  S  I  L  D  F  A  R  L  F  *
GGCCTCAATTCTGGACTTTGCCCGCTTGTTCTAAGCAGCTGCTTCCACTTTTGTATCAGA
      1690       1700       1710       1720       1730       1740

CATGCAGAGGCATGTAATGCAATAAATAAGTTGGCCTATGTAAAGAGGAGAGAGTTTGCT
      1750       1760       1770       1780       1790       1800

TTTGTCTTGTTCTAACCTTGTTTTGAACTAGTAAACTTTCAATGTAATGAGAGTTGTTAT
      1810       1820       1830       1840       1850       1860

CTTTCTA
```

FIG.3

```
          10         20         30         40         50         60
MVISKSPFIV LIFSLLLSFA LLCSGVSAYG RKQYERDPRQ QYEQCQRRCE SEATEEREQE

          70         80         90        100        110        120
QCEQRCEREY KEQQRQQEEE LQRQYQQCQG RCQEQQQGQR EQQQCQRKCW EQYKEQERGE

         130        140        150        160        170        180
                  EEPGSQF ANPAYHF
HENYHNHKKN RSEEEEGQQR NNPYYFPKRR SFQTRFRDEE GNFKILQRFA ENSPPLKGIN

         190        200        210        220        230        240
DYRLAMFEAN PNTFILPHHC DAEAIYFVTN GKGTITFVTH ENKESYNVQR GTVVSVPAGS

         250        260        270        280        290        300
TVYVVSQDNQ EKLTIAVLAL PVNSPGKYEL FFPAGNNKPE SYYGAFSYEV LETVFNTQRE

         310        320        330        340        350        360
KLEEILEEQR GQKRQQGQQG MFRKAKPEQI RAISQQATSP RHRGGERLAI NLLSQSPVYS

         370        380        390        400        410        420
NQNGRFFEAC PEDFSQFQNM DVAVSAFKLN QGAIFVPHYN SKATFVVFVT DGYGYAQMAC

         430        440        450        460        470        480
PHLSRQSQGS QSGRQDRREQ EEESEEETFG EFQQVKAPLS PGDVFVAPAG HAVTFFASKD

         490        500        510        520        530        540
QPLNAVAFGL NAQNNQRIFL AGKKNLVRQM DSEAKELSFG VPSKLVDNIF NNPDESYFMS

         550        560        570        580        590        600
FSQQRQRRDE RRGNPLASIL DFARLF
```


PCT/GB91/00914

FIG. 4A

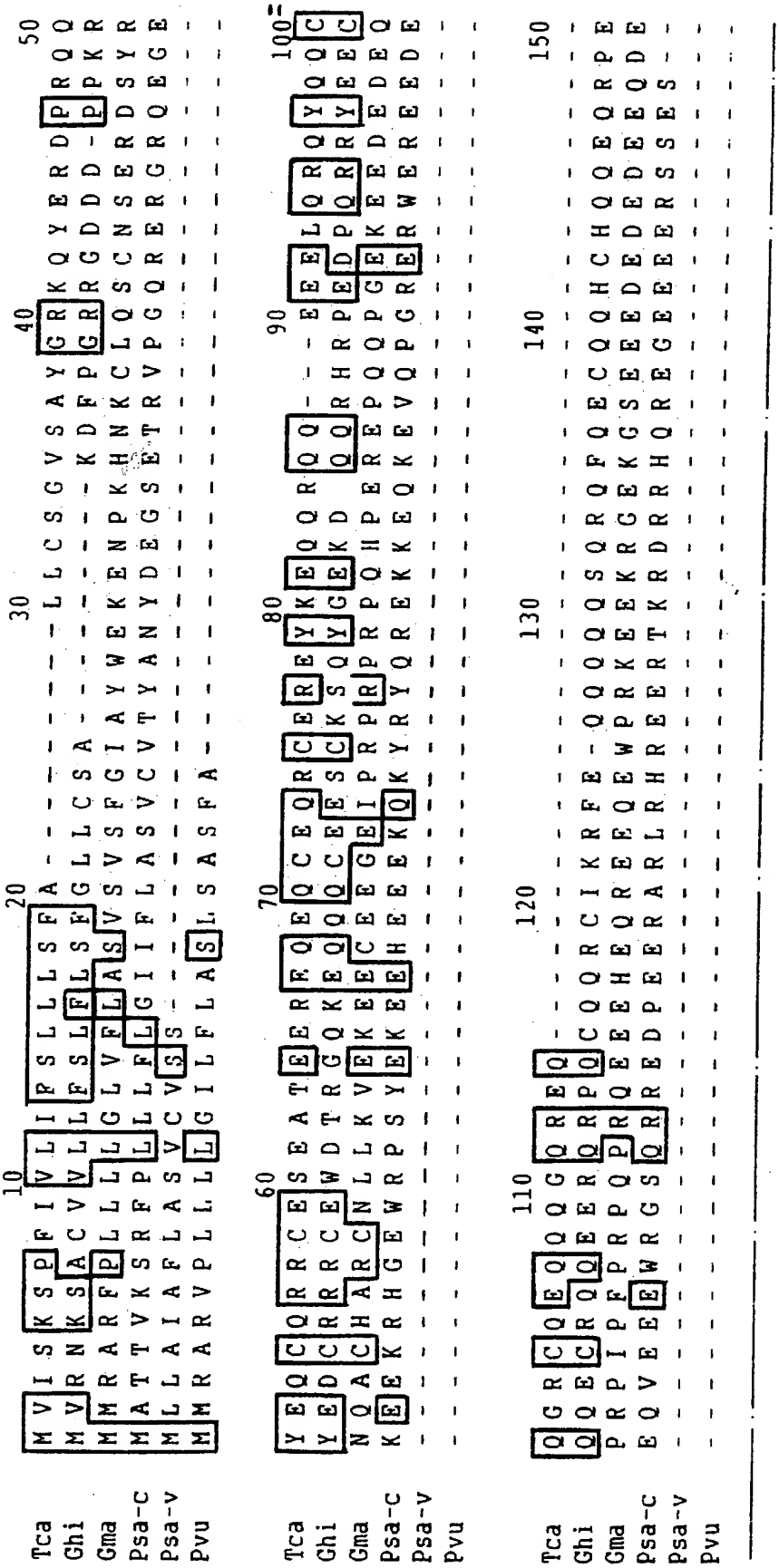


FIG. 4B

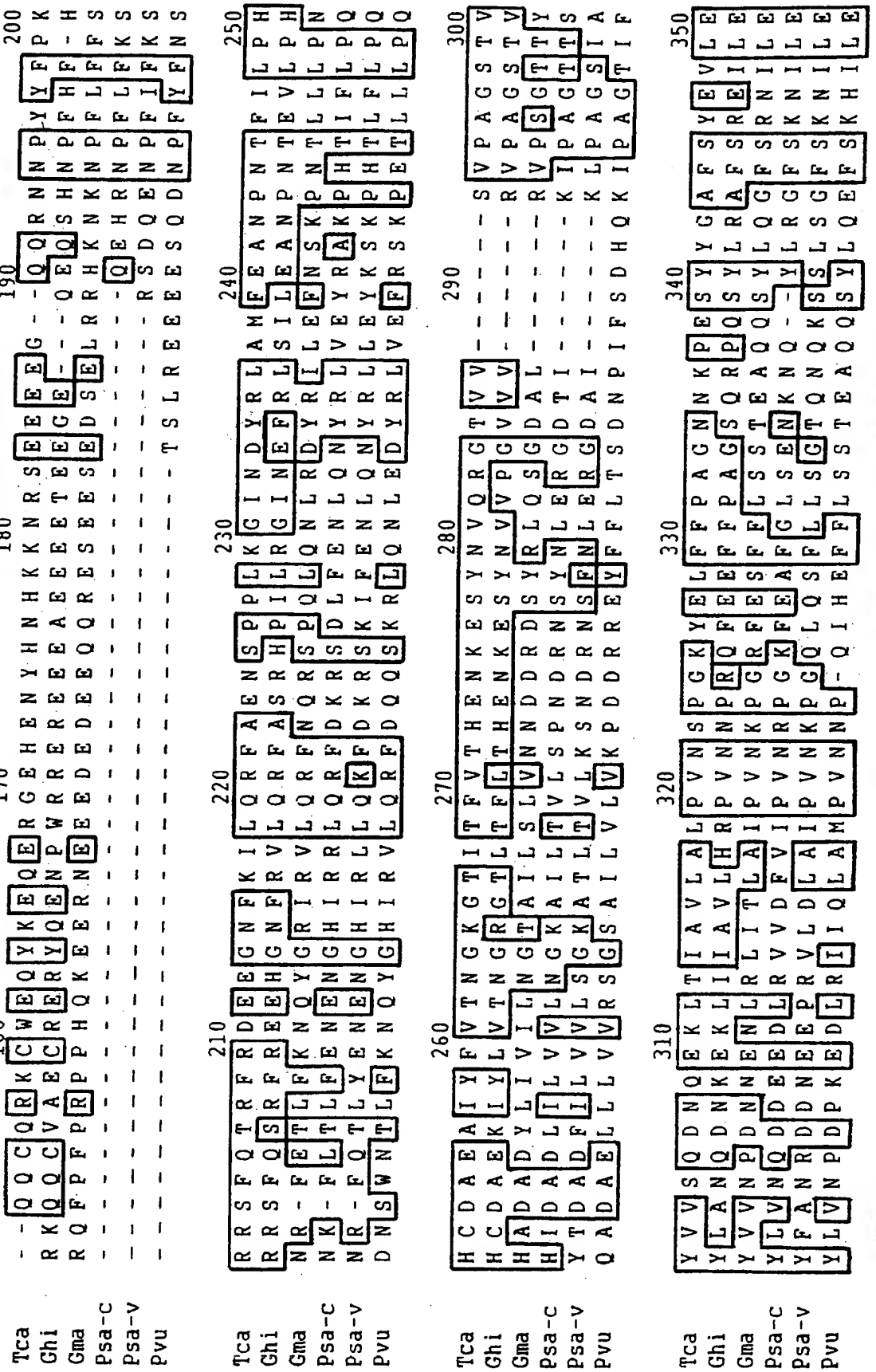


FIG. 4C

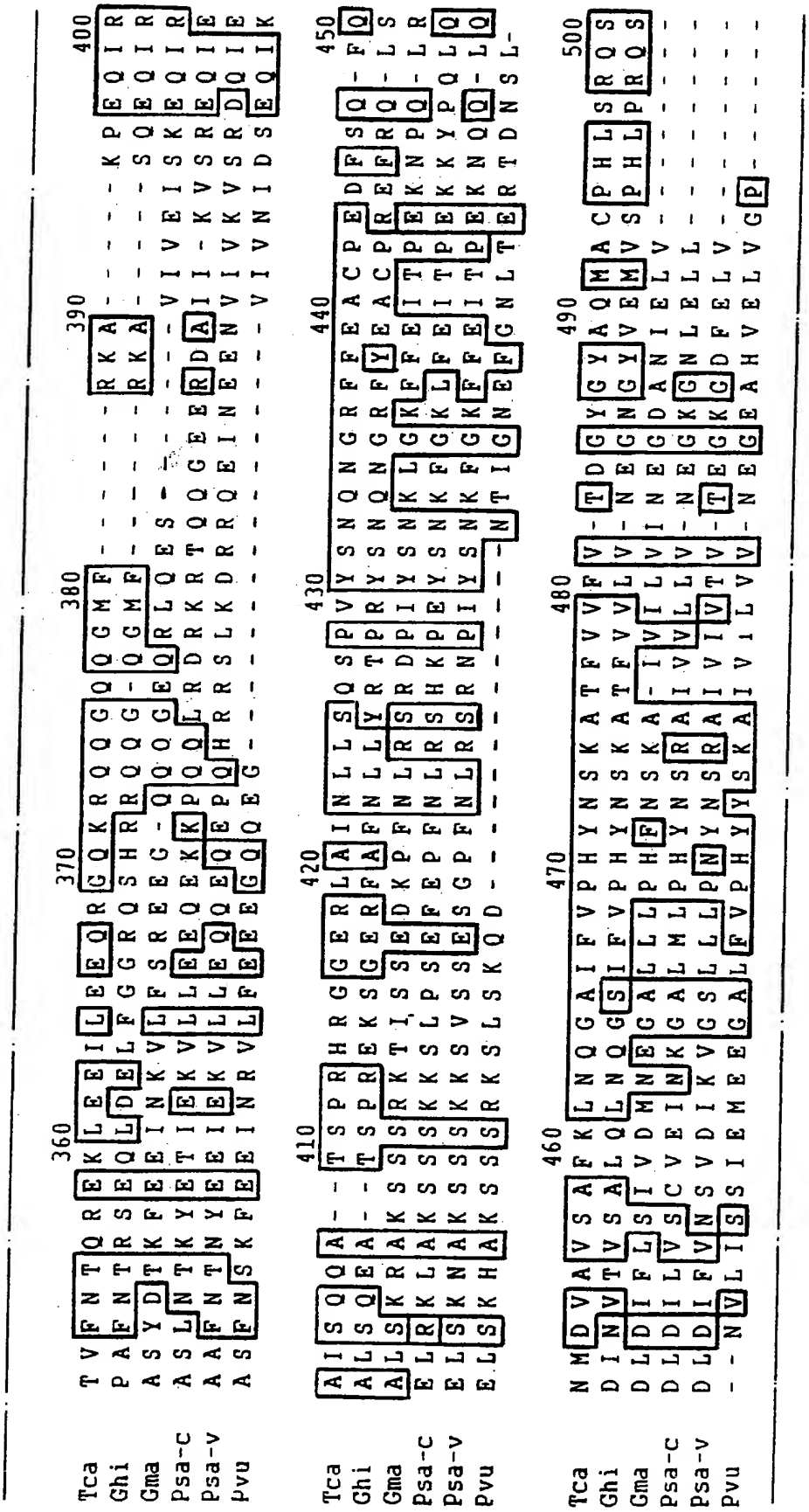
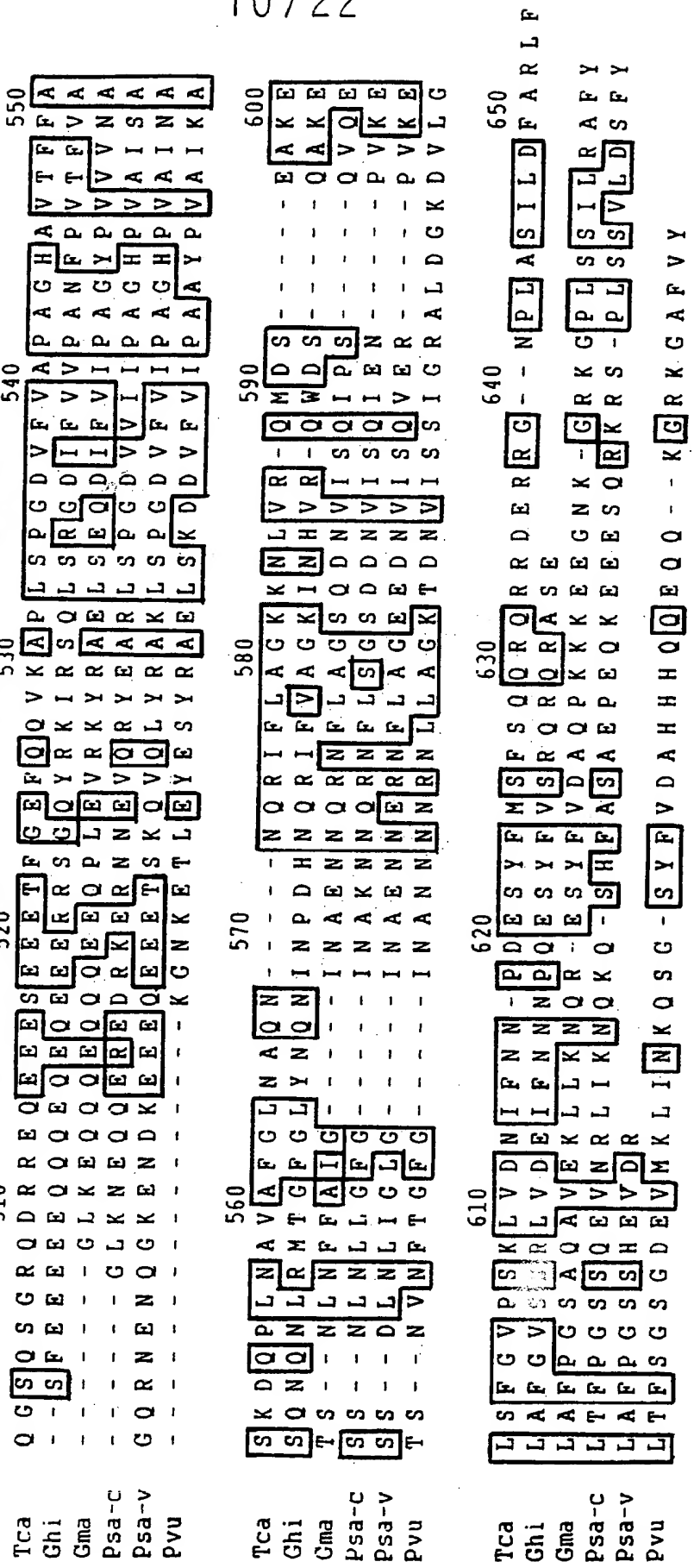
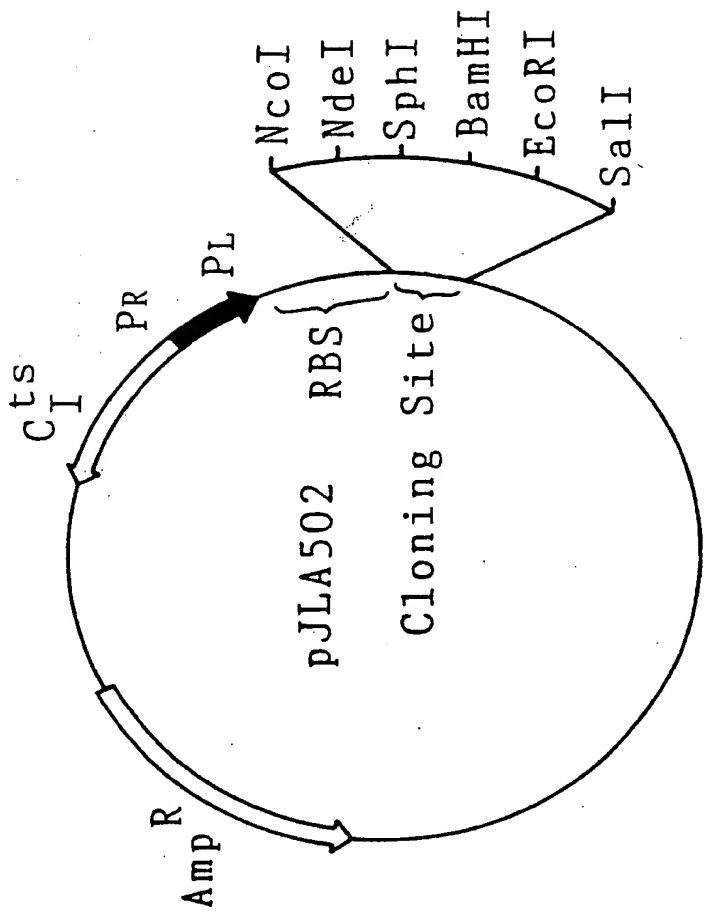


FIG. 4D





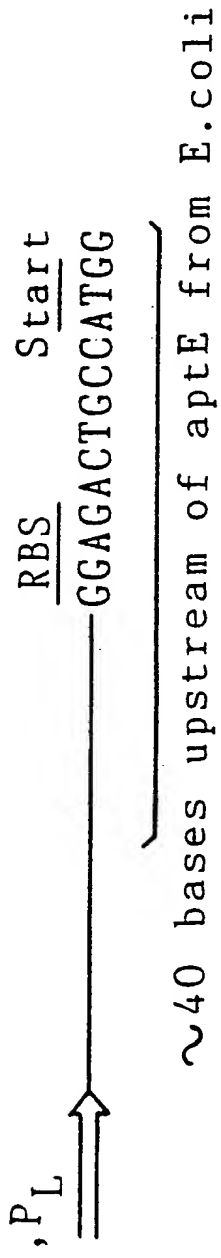
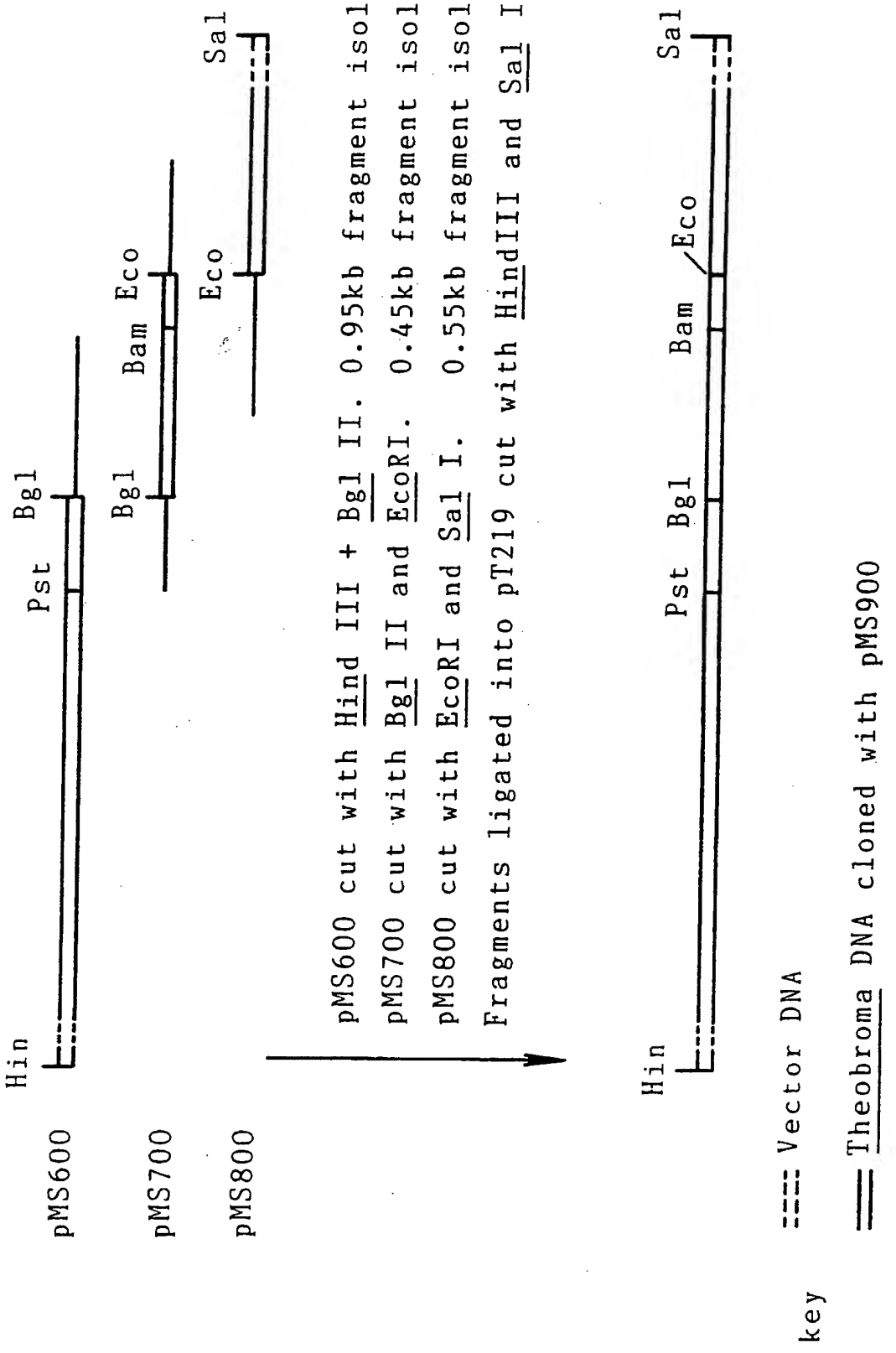


FIG. 5

FIG. 6



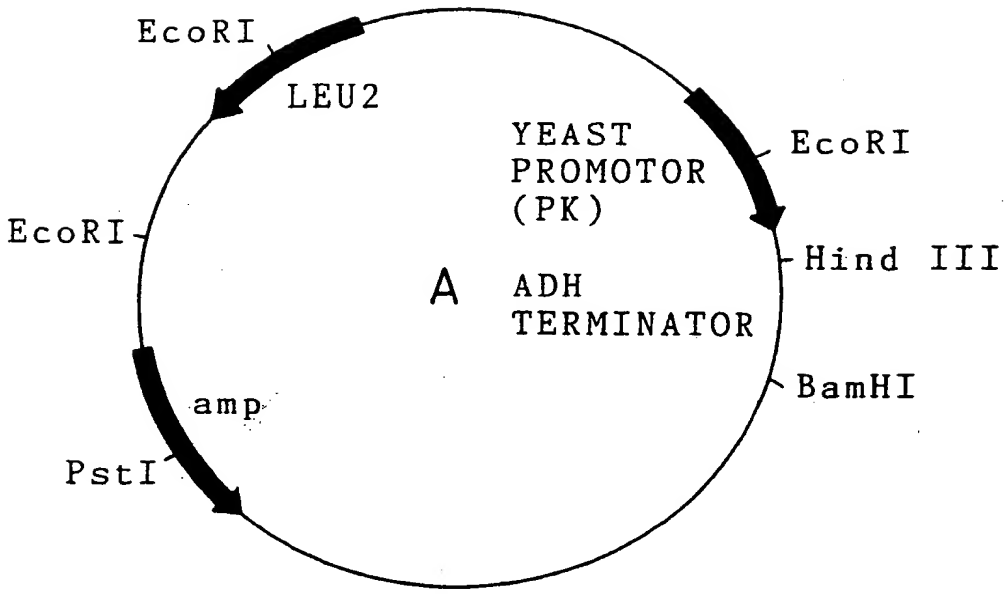


FIG. 7

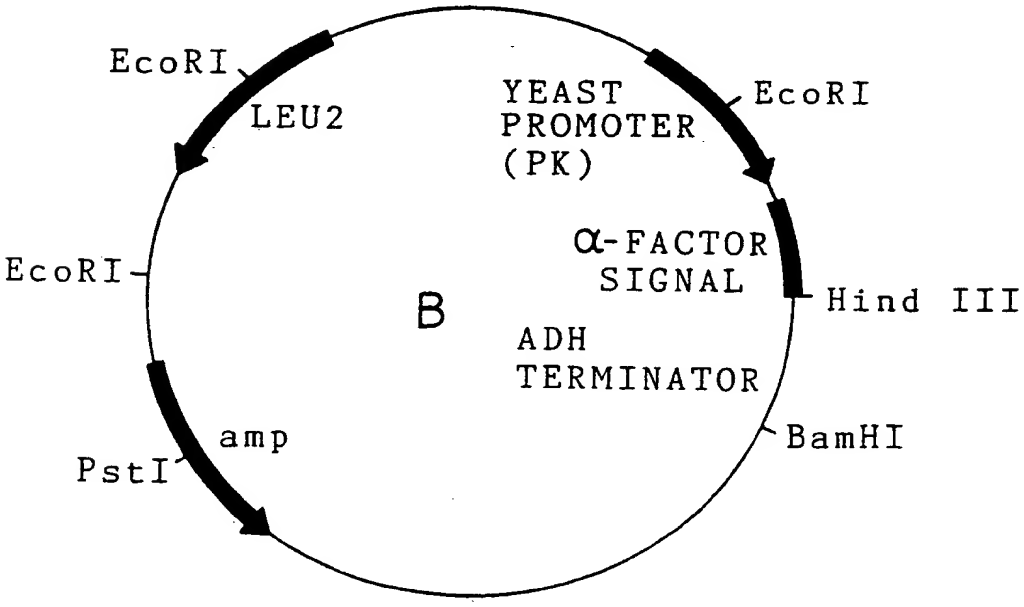


FIG. 8A

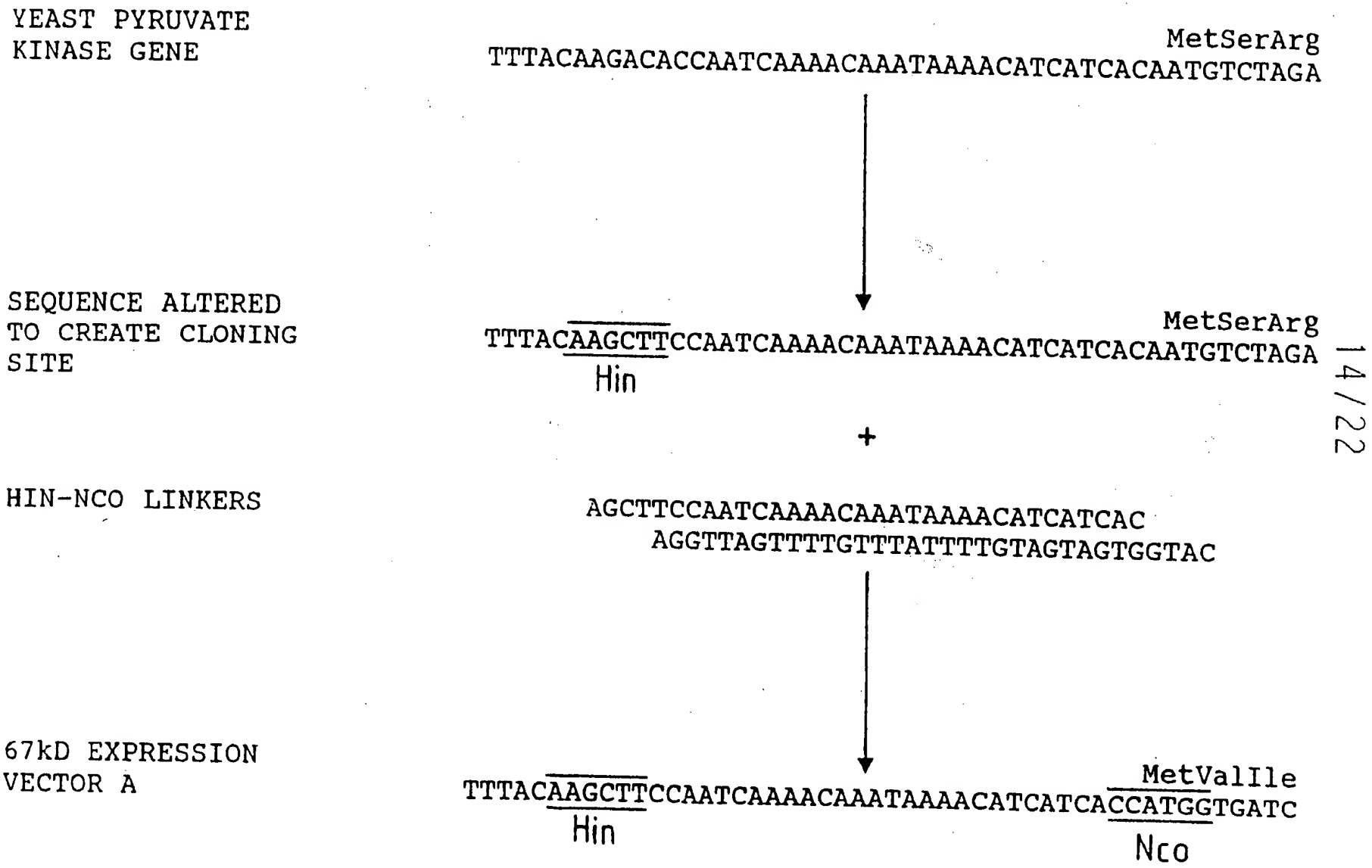


FIG. 8B

YEAST ALPHA-FACTOR
SIGNAL SEQUENCE

```
 1      231
Met----GluGlyValSerLeuAspLysArgGlu
ATG----GAAGGGGTAAGCTTGGATAAAAGAGAG
           Hin
```

HIN-NCO LINKERS

```
AGCTTGGATAAAAGAGC
    ACCTATTTTCTCGGTAC
```

IN-PHASE FUSION OF
67KD CODING REGION

```
Met---GluGlyValSerLeuAspLysArgAlaMetAlaLeu
ATG---GAAGGGGTAAGCTTGGATAAAAGAGCCATGGCGTTG
           Hin                 Nco
```

FIG. 9A

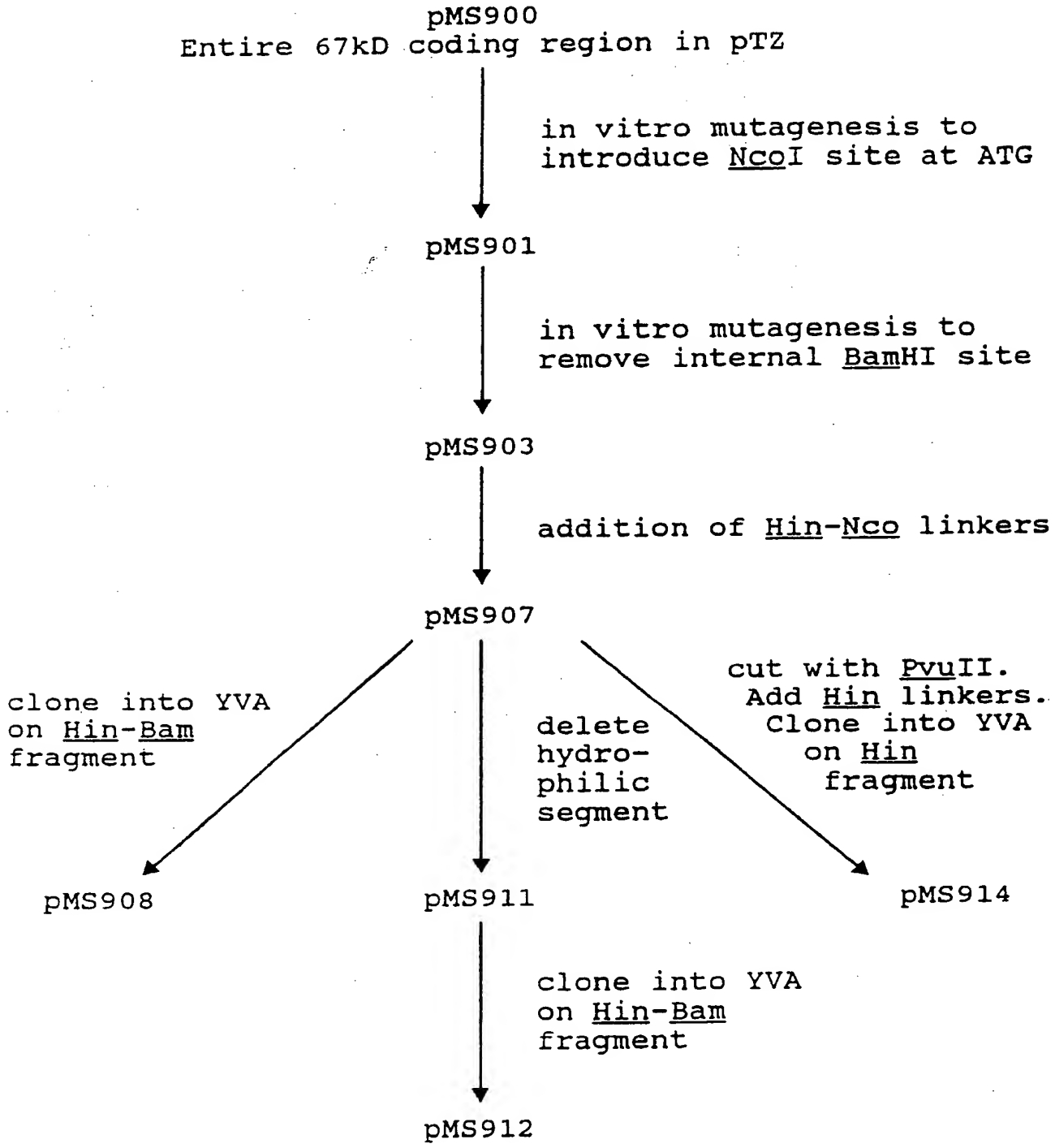
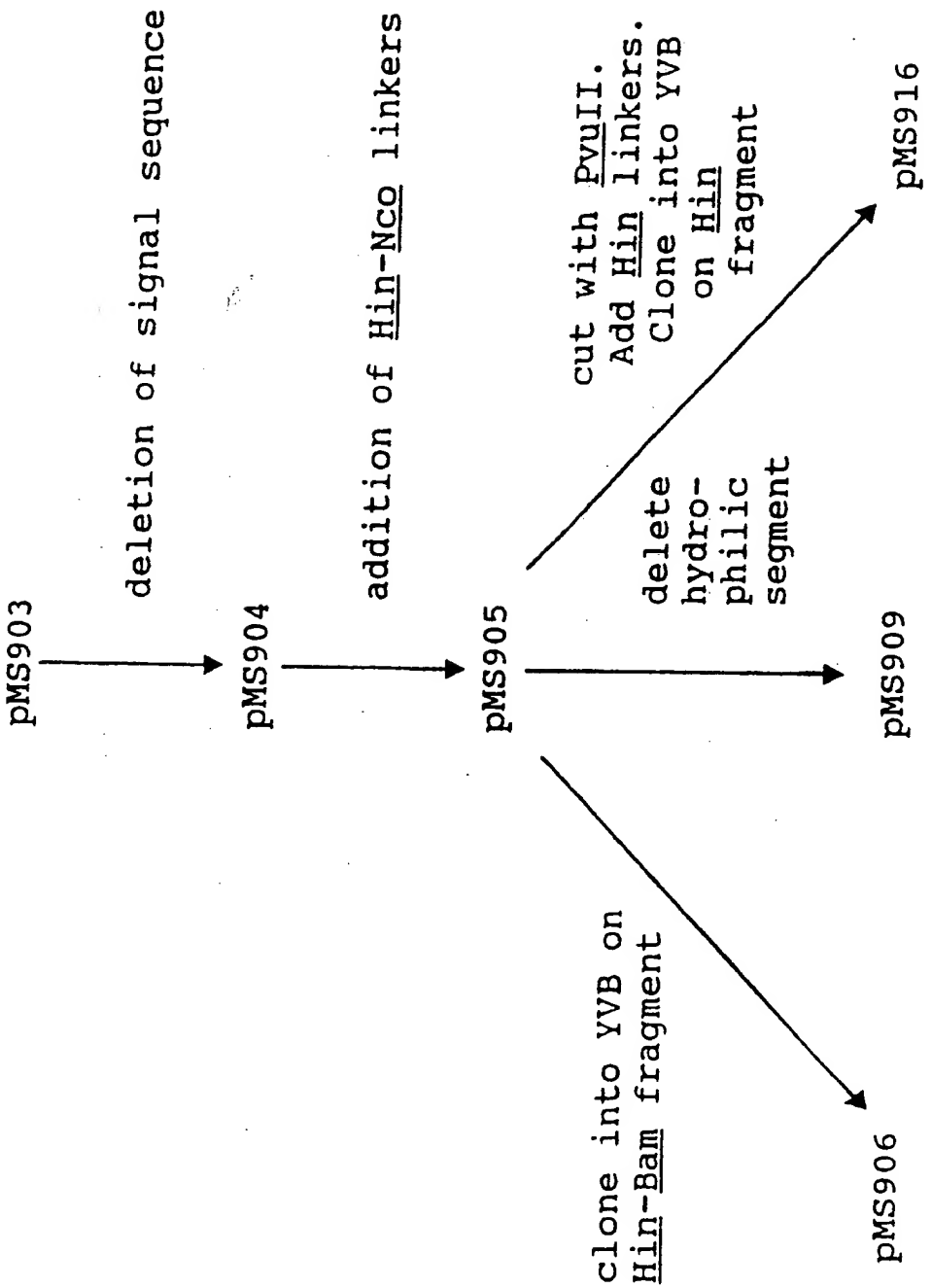


FIG. 9B



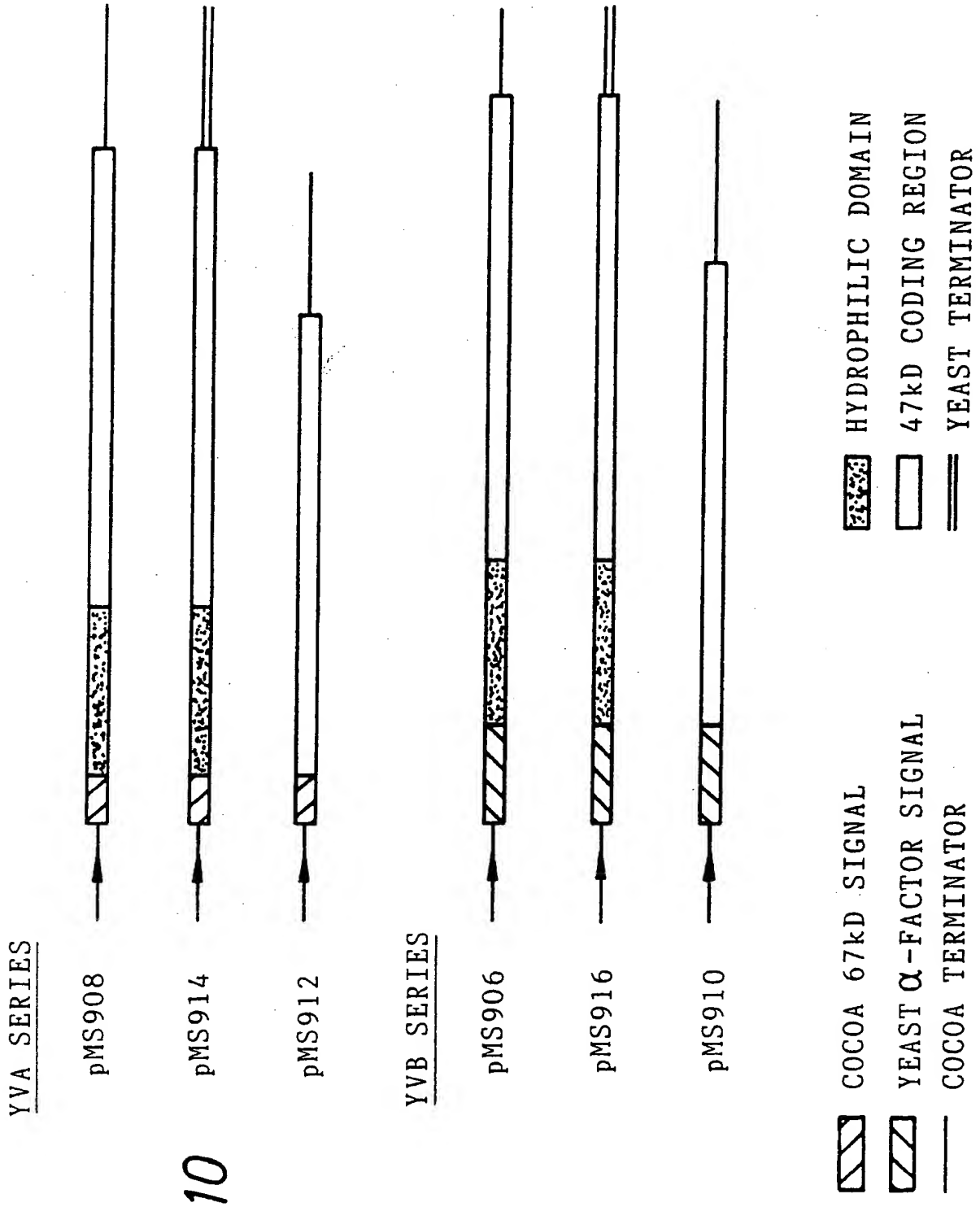


FIG. 10

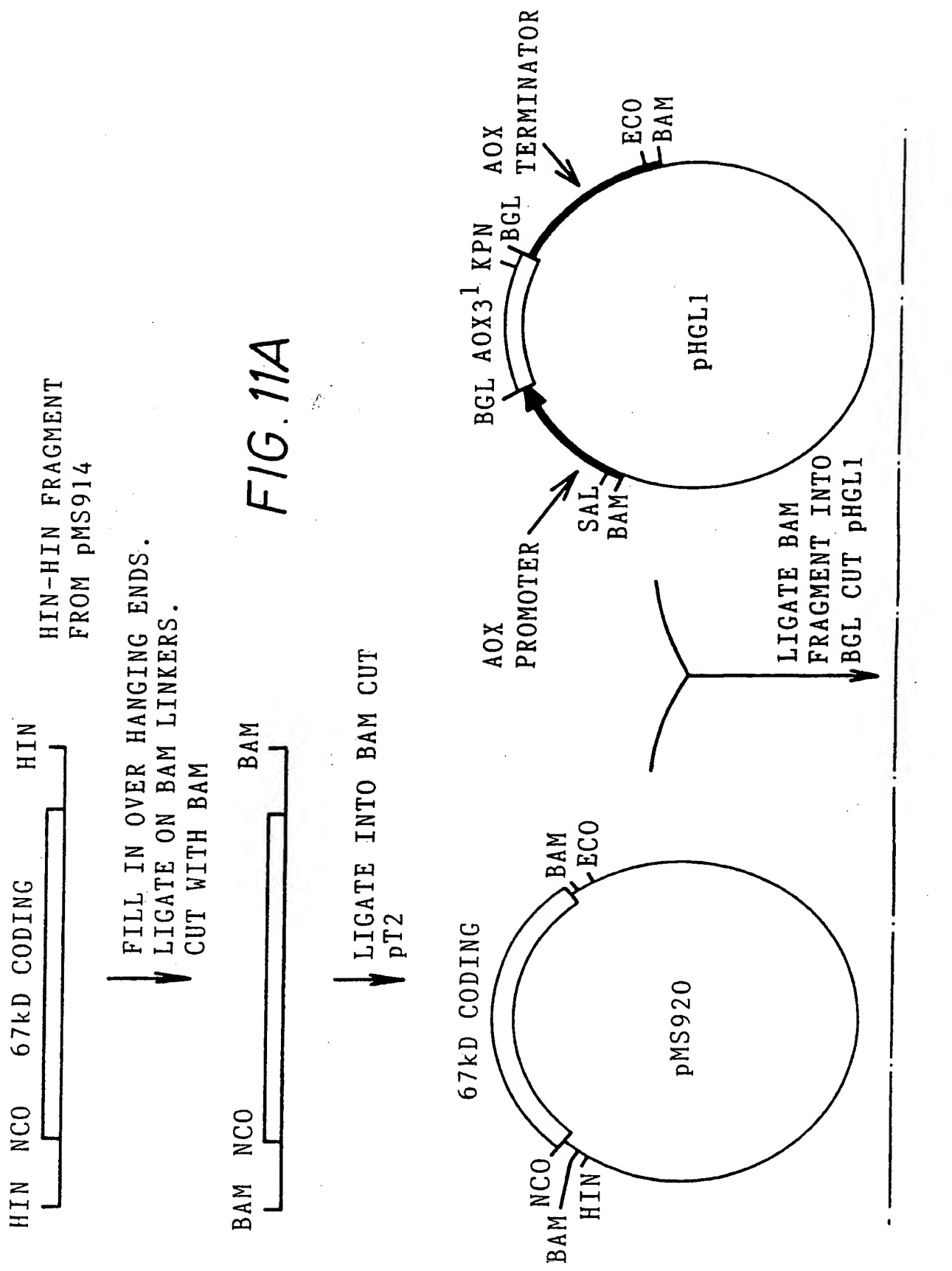


FIG. 11A

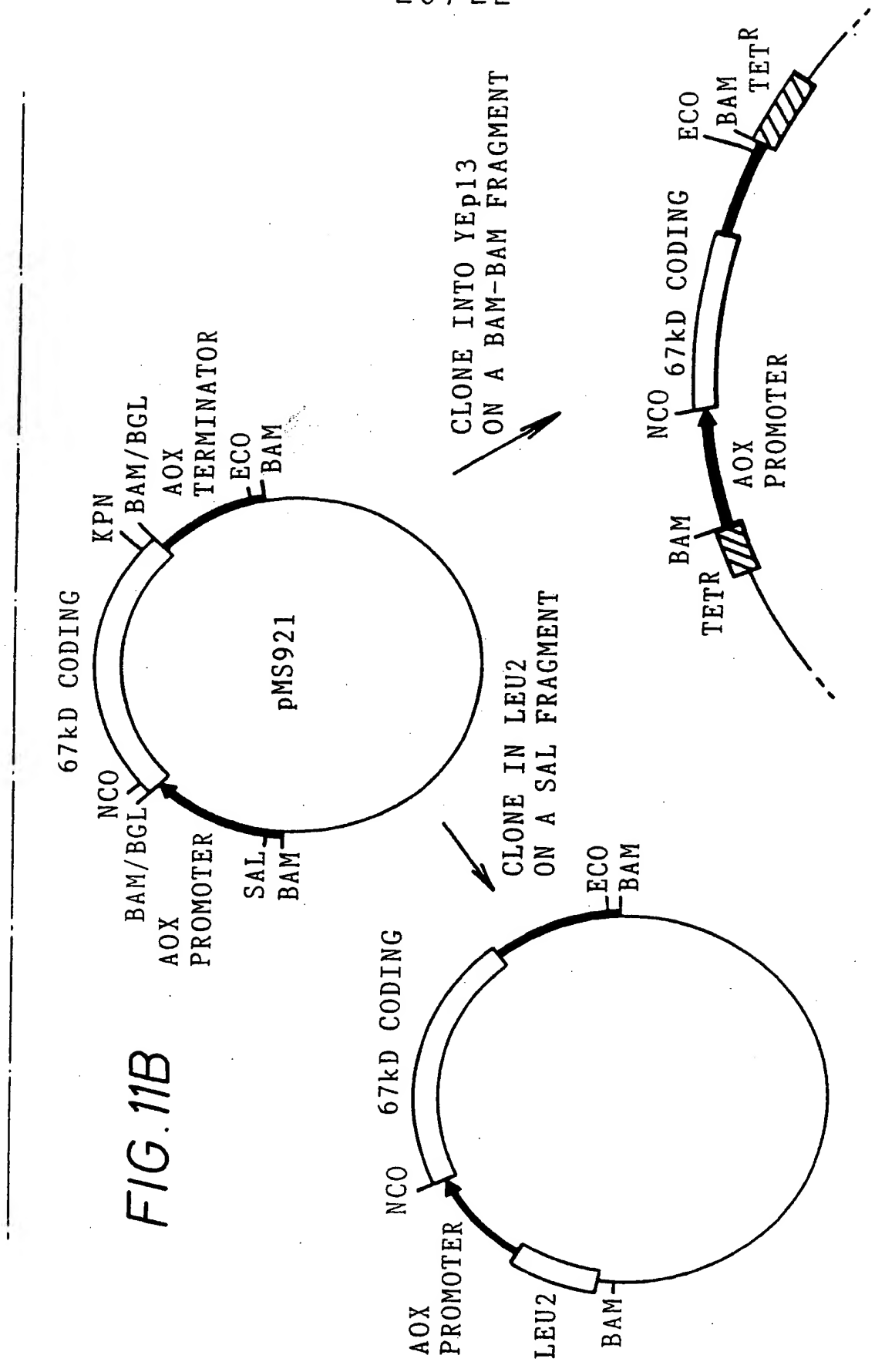


FIG. 11B

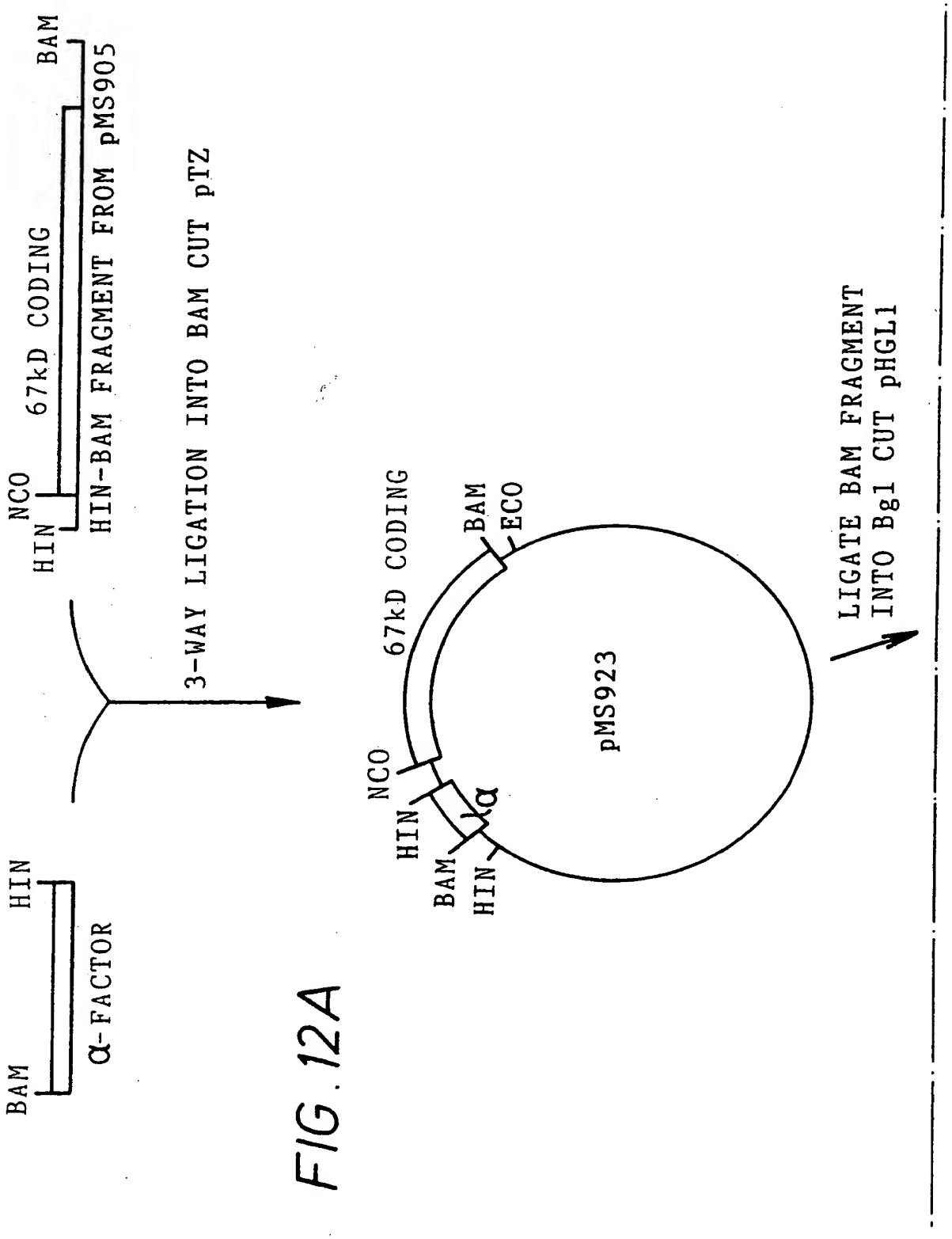
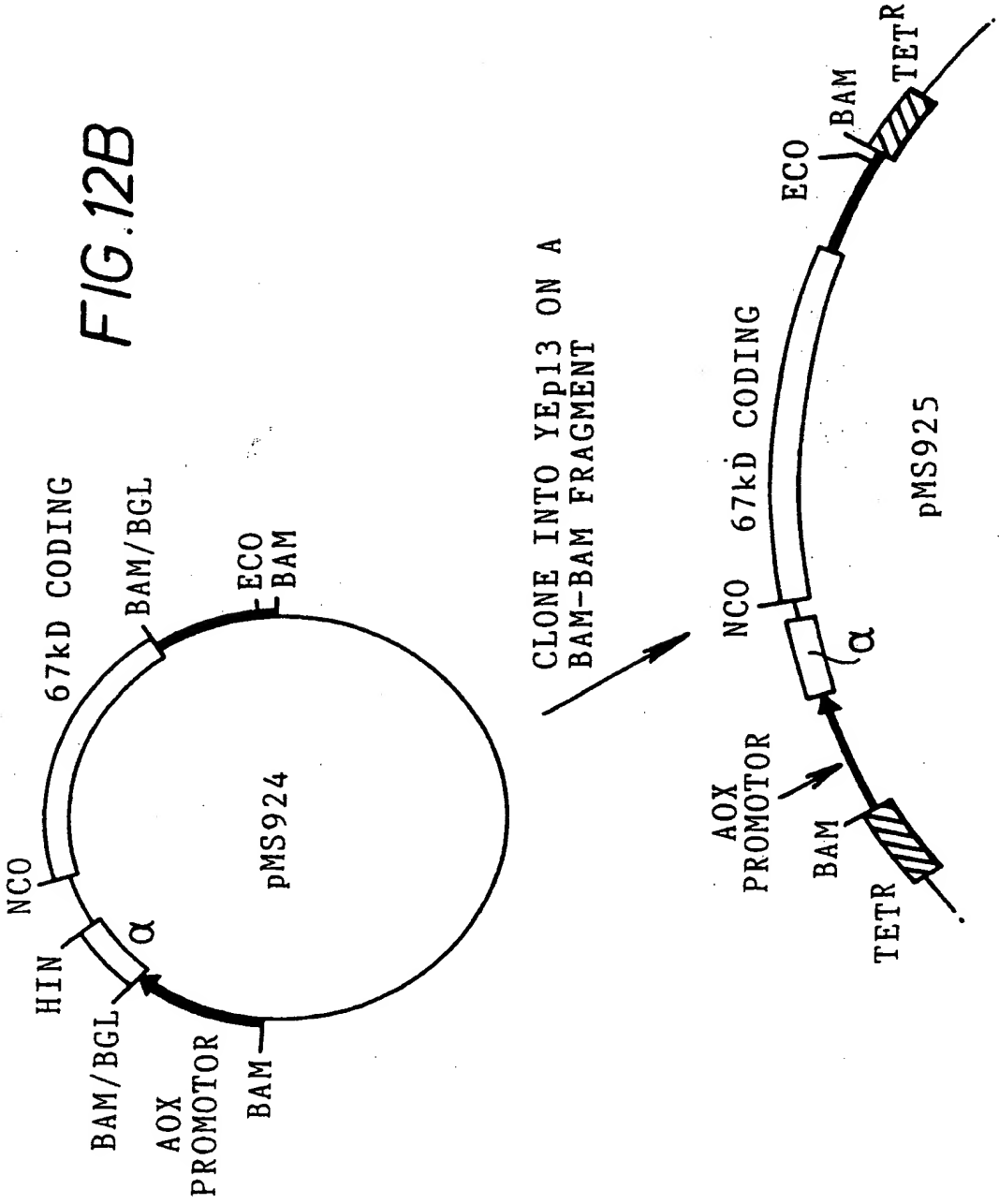


FIG. 12A

FIG.12B



INTERNATIONAL SEARCH REPORT

International Application No PCT/GB 91/00914

I. CLASSIFICATION OF SUBJECT MATTER (if several classification symbols apply, indicate all)[6]

According to International Patent Classification (IPC) or to both National Classification and IPC

Int.Cl. 5 C12N15/29 ; C07K13/00 ; C12N1/21 ; C12N1/19

II. FIELDS SEARCHED

Minimum Documentation Searched[7]

Classification System	Classification Symbols
Int.Cl. 5	C07K ; C12N

Documentation Searched other than Minimum Documentation to the Extent that such Documents are Included in the Fields Searched[8]

III. DOCUMENTS CONSIDERED TO BE RELEVANT[9]

Category °	Citation of Document,[11] with indication, where appropriate, of the relevant passages[12]	Relevant to Claim No.[13]
X	J. SCI. FOOD AGRIC. vol. 33, 1982, pages 1291 - 1304; BIEHL, B., ET AL: 'Vacuolar storage proteins of cocoa seeds and their degradation during germination and fermentation ' see the whole document	1-6
X	J. FOOD SCIENCE vol. 50, 1985, pages 946 - 950; FRITZ, P. J., ET AL: 'Cocoa seeds: Changes in protein and polysomal RNA during development ' see the whole document	1-6

-/--

° Special categories of cited documents :[10]

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier document but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art.

"&" document member of the same patent family

IV. CERTIFICATION

Date of the Actual Completion of the International Search	Date of Mailing of this International Search Report
30 SEPTEMBER 1991	20. 11. 91
International Searching Authority	**Signature of Authorized Officer**
EUROPEAN PATENT OFFICE	MADDOX A.D.

Form PCT/ISA/210 (second sheet) (January 1985)

III. DOCUMENTS CONSIDERED TO BE RELEVANT (CONTINUED FROM THE SECOND SHEET)

Category °	Citation of Document, with indication, where appropriate, of the relevant passages	Relevant to Claim No.
X	CAFE CACAO THE vol. 34, no. 1, January 1990, pages 23 - 26; PETTIPHER G. L.: 'The extraction and partial purification of cocoa storage proteins ' see the whole document	1-6
Y	---	7-20
Y	ABSTR. PAP. AM. CHEM. SOC. vol. 188, 1984, BIOL 148 WILSON, M. R.,ET.AL.: 'Cocoa theobroma -cacao seed complementary DNA library ' see the abstract 148	7-20
X	---	
	PLANT MOL. BIOL. vol. 9, no. 6, 1987, pages 533 - 546; CHLAN C. A., ET. AL.: 'Developmental biochemistry of cottonseed embryogenesis and germination. XIX. Sequences and genomic organization of the alpha-globulin, vicilin , genes of cottonseed. ' see figures 3,4	4-11,14, 15,17,20
X	---	
	CHEMICAL ABSTRACTS, vol. 109, no. 17, Columbus, Ohio, US; abstract no. 143880s, WATSON, MARTIN D. 'Isolation and expression of a pea vicilin cDNA in the yeast Saccharomyces cerevisiae ' see abstract	4-11, 14-16,19
X	---	
	PROC. NATL. ACAD. SCI. U. S. A. vol. 82, no. 2, January 1985, pages 334 - 338; CRAMER J H: 'Expression of phaseolin cDNA genes in yeast under control of natural plant DNA sequences. ' see the whole document	4-11, 14-16,19
X	---	
	CHEMICAL ABSTRACTS, vol. 106, no. 13, Columbus, Ohio, US; abstract no. 98280, CRAMER, JANE HARRIS 'Signal peptide specificity in posttranslational processing of the plant protein phaseolin in Saccharomyces cerevisiae ' see abstract	4-11, 14-16,19

	-/--	

III. DOCUMENTS CONSIDERED TO BE RELEVANT (CONTINUED FROM THE SECOND SHEET)

Category °	Citation of Document, with indication, where appropriate, of the relevant passages	Relevant to Claim No.
X	PLANT MOL BIOL vol. 11, 1988, pages 683 - 695; HIGGINS T.J.V., ET.AL.: 'The sequence of a pea vicilin gene and its expression in tansgenic tobacco plants ' see page 684 methods section ---	4-11,15, 17,19